INFORME 1082-1

MODIFICACIÓN PARA UNIFICAR LA METODOLOGÍA DE EVALUACIÓN DE LA CALIDAD DE LA IMAGEN

(Cuestión 3/11, Programa de Estudios 3A/11)

(1986-1990)

1. Introducción

La Recomendación 500 que se revisa regularmente tiene como fin instruir sobre los métodos disponibles que parecen más adecuados para la evaluación de la calidad de la imagen en un entorno de laboratorio controlado. Es necesario revisar los métodos periódicamente a fin de reflejar la evolución de los estudios en los nuevos sistemas y aplicar la evolución de la propia metodología.

Aunque los métodos descritos en los puntos 2 y 3 de la Recomendación 500 se han examinado y diseñado minuciosamente con los conocimientos disponibles, no están libres de inconvenientes. Si se conciben nuevos métodos alternativos que puedan evitar dichos inconvenientes, deben proponerse para sustituir a los métodos actuales.

Los inconvenientes principales de los métodos que figuran actualmente en los puntos 2 y 3, son los siguientes:

- Las diferencias conceptuales entre los significados de los descriptores de la escala de calidad no son necesariamente uniformes. Se sabe que varían de unos grupos lingüísticos a otros, entre grupos culturales y entre individuos, en un grado nada despreciable. El procesamiento de los resultados se basa actualmente en la aproximación de que la diferencia conceptual es uniforme; de esta manera la interpretación de resultados para indicar una medida coherente de calidad absoluta o de degradación es también una aproximación. De hecho, los resultados pueden incluso representar erróneamente las magnitudes de las diferencias hasta en un $\pm$ 50%.

- Por motivos que pueden referirse a las diferencias de significado asociadas a los descriptores mencionados, la correlación de los resultados entre laboratorios no se considera suficientemente buena para los sistemas alternativos con pequeñas degradaciones o los de gran calidad, de forma que puedan evaluarse fiablemente en los distintos laboratorios y compararse los resultados absolutos. No obstante, el orden de categorías es coherente.

- La estabilidad de los métodos mencionados en los puntos 2 y 3 de la Recomendación 500 se deriva en parte de la utilización sistemática de una referencia de gran calidad. Hay circunstancias en las que no se dispone de una referencia de gran calidad y, en dichos casos, los métodos no pueden utilizarse.

- Los métodos de doble estímulo llevan el doble de tiempo que los métodos de un solo estímulo y como consecuencia son más costosos de realizar.

Este Informe describe estudios relacionados con el desarrollo de nuevos métodos que aportan más información y que superan o dejan de lado los inconvenientes mencionados. Las áreas generales que se estudian son las siguientes:

- valoración cuantitativa (estimación de la magnitud numérica de la calidad);

- escala gráfica (evaluación de las diferencias conceptuales en los descriptores);

- escala de categorías numéricas;

- escala multidimensional;

- comparación por parejas;

- medición del umbral de visibilidad.

Para que pueda pensarse en su incorporación a la Recomendación 500, los métodos deben estar totalmente desarrollados y ofrecer ventajas significativas en comparación con los métodos recomendados actualmente.

Este Informe describe también los trabajos recientes encaminados a examinar si es posible evaluar degradaciones tales como el ruido, utilizando escalas gráficas.

2. Valoración cuantitativa

2.1 Introducción

Debido a que permite la más amplia gama y variedad de operaciones estadísticas, la valoración cuantitativa permite al experimentador no sólo establecer el orden de los elementos que se calibran, sino también describir las magnitudes relativas de cualquier atributo elegido de estos elementos. Puede así determinarse, por ejemplo, no sólo que la imagen A es mejor que la B, sino también **cuánto mejor**. Una escala de categorías permite sólo una determinación del orden de rango y no de los intervalos entre rangos. El método psicofísico más apropiado para generar escalas de relación de calidad de imagen es la **estimación de magnitudes.**

El método de la estimación de magnitudes se ha utilizado en todo el mundo desde mediados de los años 50. Las escalas dan datos de relaciones y por tanto datos de intervalos iguales. Con los datos de escalas de relación, es legítimo calcular medidas geométricas o armónicas, y también medidas aritméticas, y calcular variaciones porcentuales y también desviaciones típicas. Esta mayor precisión permite una descripción más completa de los datos. Los observadores generan sus propias escalas, evitando así un importante inconveniente de las escalas fijas, cual es el de imponer designaciones verbales y sistemas de numeración posiblemente inadecuados.

2.2 Metodología de las pruebas experimentales

2.2.1 Procedimiento

En el método de la estimación de magnitudes, los observadores producen sus propias escalas a medida que avanza el experimento. Se presenta al observador una serie de imágenes con una secuencia irregular y se le pide que asigne un valor numérico de calidad a cada imagen presentada. Las instrucciones deben incluir la siguiente información e inspirarse en este ejemplo:

"Se le presentarán una serie de imágenes aleatoriamente. Su labor consistirá en juzgar la CALIDAD de imagen de cada una asignándoles números. Evalúe la primera imagen con el número que crea apropiado. Asigne luego número a las presentaciones sucesivas proporcionales a su impresión subjetiva. No existe límite alguno para la gama de números que puede utilizar. Puede asignar números enteros, decimales o fracciones. Trate que cada número corresponda a la calidad de imagen que percibe. Por ejemplo, si una imagen le parece tres veces mejor que otra, asígnele un número tres veces mayor; si le parece sólo de un quinto de calidad, asígnele un quinto de su valor."

La gama y el número de estímulos deben ser tan grandes como sea razonable en ese experimento concreto, a fin de no limitar al observador a un pequeño conjunto de condiciones y permitirle utilizar todos los criterios disponibles para evaluar la calidad. Cada estímulo suele presentarse dos veces (con más presentaciones se obtiene poca o ninguna información adicional).

2.2.2 Evaluación del "ideal"

Para establecer una referencia que permita comparar resultados de prueba obtenidos en diferentes laboratorios, con diferentes sistemas de televisión, imágenes de prueba y así sucesivamente, debe pedirse a los observadores que, al final de cada sesión de prueba, asignen el valor numérico apropiado a su concepción de la calidad de imagen que consideren "ideal". El "ideal" tiene por objeto designar la mejor calidad de imagen posible imaginable producida por cualquier sistema de reproducción de imágenes. En el ánalisis de los datos (véase el punto 2.2.6), se normalizará la evaluación del "ideal" (es decir, de ese número) para que corresponda al número 100, valor éste que proporcionará una norma uniforme.

2.2.3 Evaluadores

Se ha comprobado que las relaciones entre las medias geométricas resultan estables con 15 observadores (naturalmente pueden ser más si se desea).

2.2.4 Imágenes de prueba

La conveniencia de las imágenes de prueba depende del experimento concreto que se realice.

2.2.5 Presentación

Las imágenes deben presentarse en un orden aleatorio, a condición de que no se presente la misma imagen (es decir, escena o secuencia de prueba) dos veces seguidas con el mismo nivel de calidad. Si es posible las imágenes presentadas a cada observador deben seguir un orden aleatorio diferente. Debe variarse la imagen de estímulo inicial para cada observador, pero no es necesario variar el nivel de calidad. Es aconsejable iniciar cada secuencia en algún punto medio de la gama, y no en ninguno de los extremos.

Una sesión de observación debe durar aproximadamente media hora, incluidas las explicaciones y los preparativos. La secuencia de prueba puede iniciarse con unas pocas imágenes indicativas de la gama de calidad de imagen (aunque esto no es necesario ni tampoco debe decirse en ningún caso al observador cuál puede ser esa gama). Las apreciaciones de estas presentaciones preliminares no deben tenerse en cuenta en los resultados finales.

2.2.6 Normalización y cálculo de la media de las estimaciones de magnitud

La medida apropiada y más comúnmente empleada de la tendencia central con datos de estimación de magnitud es la media geométrica, que tiene en cuenta la distribución de las respuestas y presenta la ventaja de evitar una evaluación extrema que influya excesivamente en el resultado. Da una estimación sin sesgo del valor esperado de los logaritmos de las estimaciones de magnitud. Pese a los diferentes números que los observadores pueden haber asignado al primer estímulo, no es necesaria ninguna normalización antes de calcular la media. Las relaciones entre las medias geométricas no son afectadas aunque los observadores utilicen diferentes unidades para sus escalas subjetivas. Sin embargo, la normalización resultará necesaria para ciertas operaciones estadísticas subsiguientes y para la comparación entre laboratorios. Para conseguirla, deben efectuarse para cada dato de observador los siguientes cálculos de normalización con respecto al «ideal».

Las medias geométricas se normalizarán para que al «ideal» le corresponda el valor normalizado 100. Para lograrlo, debe calcularse la media geométrica de las respuestas numéricas. Todas las medias geométricas se multiplican entonces por el factor común $100/R_i$ (donde R_i es el valor numérico de la respuesta «ideal»). Este sencillo procedimiento sirve para definir el «ideal» como 100, y al mismo tiempo para ajustar proporcionalmente las respuestas medias a todos los demás estímulos.

2.3 Estudios sobre la validez de la valoración cuantitativa

2.3.1 Comparación del empleo de una escala de valoración por categorías con un solo estímulo y una valoración cuantitativa

2.3.1.1 Introducción

Se realizaron un par de experimentos de calidad de imagen [CCIR, 1982-86a] siguiendo la escala de calidad de cinco notas de la Recomendación 500 y, a título comparativo los del método de la estimación de magnitudes. El objeto era examinar el efecto del contexto sobre las dos metodologías de prueba. La razón de tal interés en este momento es la aparición de imágenes de televisión muy mejoradas y su efecto de ampliación de la gama de calidades de imagen.

2.3.1.2 Procedimiento

Método de prueba

La prueba se hizo individualmente con cada observador y se utilizaron dos métodos de evaluación de la calidad. Para la evaluación de categoría se utilizó la escala de calidad del CCIR como método de estimación de la calidad de las imágenes de prueba. Se calculó la puntuación para cada imagen de prueba. La evaluación por estimación de magnitudes se hizo según los procedimientos descritos en el punto 2.2 de este Informe.

Aparatos y disposición

Los observadores se sentaban a una distancia de observación de tres veces la altura de la imagen. Ambas pantallas de rayos catódicos funcionaban con una frecuencia de trama de 60 Hz y tenían diagonales de 19 pulgadas (48,3 cm).

Había cuatro niveles de calidad de imagen en una prueba y cinco en la otra: A, B, C y D, que corresponden respectivamente a los sistemas NTSC de 525 líneas con decodificador por filtro de ranura (A), NTSC de 525 líneas con decodificador por filtro de peine (B), las 3 señales de color RGB obtenidas directamente de la cámara (C), e imágenes de alta definición (D). Todas estas imágenes se juzgaron de calidad buena a excelente (gama estrecha). Los niveles de estímulo X e Y en la prueba de gama ampliada eran los del sistema NTSC de 525 líneas, con filtro de ranura y ruido añadido. Los niveles de S/N para esta prueba de gama ampliada eran de 22 (X) y 32 (Y) dB. En esta prueba había menos puntos de datos RGB, pues cada observador calibraba RGB sólo una vez, como la última evaluación de cada sesión de prueba.

Entre las dos pruebas mediaron aproximadamente cinco semanas.

Evaluadores

Los 67 evaluadores tenían una agudeza y una visión cromática normales sin o con corrección previa. Ningún observador había participado antes en un experimento de estimación de magnitudes, pero algunos tenían experiencia con escalas de cinco notas.

Participaron tres grupos de evaluadores:

Un grupo de 9 observadores entre los expertos del laboratorio. Se trataba de hombres que trabajan en el campo de la ingeniería de televisión. Sus edades variaban de 27 a 65 años.

Un grupo de 47 observadores no expertos entre el personal de laboratorio. Se trataba de mujeres y hombres cuyas edades variaban de 33 a 60 años.

Un tercer grupo de 11 estudiantes de segunda enseñanza, que se formó para equilibrar el conjunto. Eran estudiantes de ambos sexos, de 16 años de edad.

Condiciones de observación

Las condiciones de observación cumplían en general la Recomendación 500 exceptuando la distancia de observación que era de tres veces la altura de la imagen.

Imágenes de prueba

Las tres imágenes de prueba eran transparencias de color, de 8 × 10 pulgadas, (8 × 25,4 cm) de la región de Stamford, en el estado de Connecticut, Estados Unidos de América. Se iluminaba mediante una Porta-Pattern. Estas imágenes se seleccionaban específicamente para asegurar la máxima calidad posible y un volumen razonable de información de frecuencia espacial elevada.

Se enfocaban dos cámaras sobre el Porta-Pattern: una cámara de 525 líneas de calidad media y una cámara de televisión de alta definición de 1125 líneas. Todas las imágenes se aplicaban desde las cámaras a codificadores cuando así convenía, o bien directamente a las unidades de visualización.

2.3.1.3 Resultados

Pruebas con la escala de cinco notas

Los tres niveles de calidad que eran comunes a ambas pruebas estaban algo desplazados. La imagen tratada con un filtro de ranura era la más desplazada (de 1,88 a 3,56), con un desplazamiento del 63% de la gama total de respuesta mínima a máxima. En otras palabras, el desplazamiento de la evaluación entre el experimento de gama estrecha y el experimento de gama ancha fue casi igual que la gama total de valoraciones utilizadas en el experimento de gama estrecha.

Es de señalar también la evidente carencia de significado de las designaciones verbales: la imagen que se había valorado como «mediocre» pasaba a ser «buena».

Pruebas con el método de estimación de magnitudes

Aparecen aquí desplazamientos similares, pero son bastante pequeños comparados con los de la escala de cinco notas. Una vez más la imagen tratada con un filtro de trampa es la más desplazada, pero en este caso el desplazamiento análogo es sólo del 34% (frente al 63%).

Otro aspecto de interés es que cuando se amplía la gama de estímulo, se amplían los números dados por los observadores. La gama total de números utilizados en la prueba de gama estrecha fue de 43 (19,5 a 62,5) en tanto que en la prueba de gama ampliada fue superior a 60 (4,12 a 64,5). Esta es otra prueba de que la escala de magnitudes es más adecuada para la tarea y más naturalmente adaptable. Los observadores se comportaron más apropiadamente ampliando sus escalas en respuesta a una mayor gama de calidad de estímulo. Con la escala de calidad, cuando se ampliaba la gama de estímulos, los números y las designaciones permanecieron fijos.

Por último, es interesante señalar que los resultados de los relativamente "ingenuos" estudiantes de segunda enseñanza mostraron que el efecto de la gama de estímulo era muy pequeño, es decir, que había muy poco desplazamiento entre las pruebas de gama estrecha y de gama ancha.

2.3.1.4 Conclusiones

Las valoraciones cuantitativas son mucho menos afectadas por los cambios en la gama de los estímulos que las escalas de categorías de cinco notas.

Las valoraciones cuantitativas hacen innecesaria la interpretación lingüística, pero exigen que el evaluador tenga una idea de las proporciones.

Las valoraciones cuantitativas tienen la virtud de introducir intervalos y relaciones significativos en las respuestas numéricas, proporcionando así información adicional acerca de cuánto supera una imagen a otra.

2.3.2 Comparación de la utilización de una escala de calidad mediante el procedimiento de doble estímulo y una valoración cuantitativa

2.3.2.1 Introducción

Otro par de experimentos relativos a la calidad de la imagen se realizó en Francia (CCIR, 1986-90b) utilizando la escala de calidad con doble estímulo y un método de valoración cuantitativa. El interés de tales experimentos deriva de las mismas razones indicadas en el punto 2.3, pero en este caso se emplea el método de evaluación continua de calidad con doble estímulo en lugar del método simple con la escla de apreciación de calidad.

2.3.2.2 Procedimiento

Método de prueba

Los métodos de prueba fueron los mismos que se indican en el punto 2.3.1, con excepción de las siguientes diferencias:

- los evaluadores efectuaron las pruebas en grupos de 4;
- se utilizó el mismo orden de presentación para todos los evaluadores;

- el procedimiento de doble estímulo se ajustó a la segunda variante del punto 3 de iniciación de la Recomendación 500;

- durante el periodo de instrucción se mostraron algunos ejemplos de imágenes.

Aparatos y disposición

Los evaluadores estaban sentados a una distancia de observación de 6 H. Se empleó una frecuencia de trama de 50 Hz: las pantallas tenían diagonales de 51 cm.

Se emplearon dos gamas de calidades (o degradaciones) que fueron evaluadas con los diversos códecs que se indican a continuación:

i) reducida; RGB, MICD1, MICD2 y SECAM;

ii) Amplia; RGB, MICD1, SECAM, MICD1 + ruido, SECAM + ruido.

Evaluadores

En cada prueba intervinieron 15 evaluadores como mínimo. Cada uno de ellos sólo participó en una prueba. Se trataba de personas no expertas.

Condiciones de observación

Las condiciones de observación se ajustaban a la Recomendación 500.

Imágenes de prueba

Se utilizaron cuatro diapositivas de prueba de la UER.

2.3.2.3 Resultados

Se probaron cuatro parámetros:

- tres tipos de estabilidad:

  - intragrupo: se empleó dos veces el mismo grupo para el mismo experimento;

  - intergrupo: se efectuó una comparación de los resultados de dos grupos diferentes;

  - efecto de contexto (gama): se efectuó una comparación de las dos gamas.

- sensibilidad: una comparación del orden atribuido a las pequeñas degradaciones.

Se empleó la prueba de la "t" de Student a fin de determinar la significación de las diferencias.

ARCHIVES U.I.T. GENÈVE

Todos los experimentos se sometieron a dos análisis para obtener resultados absolutos y relativos. En realidad, se efectúan frecuentemente dos tipos de evaluaciones: la evaluación de la degradación con respecto a una referencia y la evaluación de la calidad absoluta. Por consiguiente, para mostrar el comportamiento de cada método en ambos casos, se suministran los resultados del tratamiento de las calificaciones directas y de las diferencias de calificación entre la referencia y el objeto de la prueba.

Los resultados de las pruebas intragrupo muestran una estabilidad aceptable; la "t" permanece dentro del intervalo de confianza del 90% ($t = 1,7$) y las desviaciones típicas son similares.

De esta prueba depende la posibilidad de comparar las mediciones de distintos laboratorios. Los dos procedimientos son equivalentes, ya que los valores de "t" están próximos al valor antes indicado (1,7) o son inferiores a éste. Es necesario proseguir los estudios en diferentes laboratorios para investigar la estabilidad intergrupal de las apreciaciones utilizando estos métodos.

De los resultados de las pruebas de gama pequeña y amplia se advierte que en el caso de los resultados absolutos, el valor de "t" rebasa considerablemente el intervalo de confianza del 90%. En cambio, en los resultados relativos, el método de estimación de magnitudes puede proporcionar un valor de "t" bajo, lo que indica que esta evaluación es bastante estable. Esta diferencia entre el tratamiento absoluto y el relativo muestra que, en la estimación de magnitudes, las diferencias en los resultados derivadas de la variación de la gama de degradaciones sólo se deben a un deslizamiento global.

Por último se analizó la sensibilidad de los dos métodos mediante la comparación de la clasificación de las degradaciones que están próximas en magnitud en cada uno de los experimentos. Los dos métodos parecen igualmente aptos para proporcionar una clasificación de las pequeñas degradaciones, pero esta clasificación difiere, ya que el criterio aplicado por los evaluadores aparentemente no es el mismo en los dos procedimientos. El procedimiento de doble estímulo parece inducir a un análisis local y en este caso SECAM es el superior. La escala de relaciones, en cambio, induce aparentemente a un análisis global que hace preferibles las degradaciones digitales (MICD1 y MICD2).

2.3.2.4 Conclusión

De este estudio de los procedimientos se derivan las siguientes conclusiones:

- La utilización práctica de un método de valoración cuantitativa modificado es conveniente para la evaluación subjetiva ordinaria de las imágenes de televisión.

- Ningún procedimiento puede suministrar calificaciones absolutas fiables sin ninguna referencia.

- Cuando la gama de degradaciones es la misma para cada prueba, tanto el método del doble estímulo como el de valoración cuantitativa son suficientemente estables para permitir la comparación de los resultados procedentes de laboratorios diferentes.

- En el caso de gamas de degradaciones diferentes, sólo el procedimiento de valoración cuantitativa es suficientemente estable para suministrar resultados relativos fiables que pueden compararse de un experimento a otro, y ello sólo si se incluye en la prueba una referencia implícita idéntica. Por consiguiente, se necesita una referencia para cada tipo de evaluación subjetiva: televisión ordinaria, alta definición, etc.

- Los criterios conforme a los cuales los evaluadores se forman su opinión pueden no ser idénticos para los dos métodos. El procedimiento basado en la valoración cuantitativa parece ser más apto para las evaluaciones globales de la calidad subjetiva de la imagen.

3. Escala gráfica

3.1 Introducción

Las escalas gráficas se han empleado para determinar los intervalos percibidos en relación con términos descriptivos. Se han aplicado escalas a adjetivos y adverbios a fin de determinar su fuerza relativa como modificadores de sustantivos y verbos. La escala de calidad del CCIR (Recomendación 500) consta de cinco (5) términos cualitativos a los que se ha aplicado una escala para determinar los intervalos correspondientes en inglés (EE.UU.), francés, alemán e italiano. Los resultados fueron sorprendentemente similares en las separaciones del intervalo (Jones y McManus, 1986).

Los resultados de las pruebas de escalas gráficas poseen un valor intrínseco. Los evaluadores formulan juicios en sus propios idiomas, sin estar sujetos a limitación alguna ni a la necesidad de una interpretación numérica. Estas escalas resultantes se han utilizado para probar la calidad de la imagen pidiendo a los evaluadores que indiquen con una marca en la línea el lugar en que, a su juicio, se sitúa la calidad de la imagen en la escala (Jones, 1986).

El Grupo Interino de Trabajo 11/4 opina que este método de evaluación subjetiva, debido a su extrema simplicidad y facilidad de empleo, puede convertirse en un método de prueba útil en el ámbito internacional. Se tiene la esperanza de que otras administraciones repetirán estos estudios en sus propios idiomas, aplicando las instrucciones y orientaciones que se proporcionan a continuación.

3.2 Metodología de las pruebas experimentales

3.2.1 Procedimiento

3.2.1.1 Prepárense hojas con largas líneas verticales (en el estudio original se utilizó una línea de 18 cm en un papel de 8 x 11 pulgadas) marcando el principio y el final de éstas. En una casilla situada en el ángulo escríbase una de las palabras objeto de la prueba (una palabra por página).

3.2.1.2 Dispónganse las hojas en el mayor número posible de órdenes aleatorios diferentes.

3.2.1.3 Preséntense a cada evaluador un conjunto de hojas. Ha de pedírsele que marque en la línea, en cada hoja, dónde estima que se sitúa el significado de la palabra escrita en el ángulo, con respecto a los dos extremos. Debe repetirse la misma operación en todas las hojas; no debe imponerse ningún límite de tiempo y debe permitirse al evaluador mirar las hojas anteriores o posteriores. No ha de darse ninguna otra instrucción o explicación, excepto un ejemplo o una repetición de la instrucción antes indicada. El experimentador no debe indicar el lugar en que ha de situarse ninguno de los términos. Tampoco debe influir en el evaluador o ayudarlo una vez comenzada la sesión. En el estudio original pocas personas tuvieron dificultad en comprender la tarea. Cuando esto ocurre, a menudo lo más conveniente es elegir otro evaluador.

3.2.2 Evaluadores

Han de recogerse respuestas del mayor número de evaluadores posible (por ejemplo, >20/grupo) y del mayor número posible de zonas comprendidas en la región lingüística. Los resultados de cada grupo deben compararse en primer lugar para determinar las diferencias percibidas dentro de las regiones lingüísticas del país.

3.2.3 Promediación de los resultados de la escala gráfica

Puede asignarse un valor a cada respuesta midiendo la distancia de un extremo de la línea a la marca efectuada por el evaluador. Seguidamente pueden calcularse y representarse gráficamente las medias geométricas o aritmética y las desviaciones típicas.

3.3 Resultados de las evaluaciones efectuadas hasta la fecha sobre los intervalos de percepción entre descriptores

El Documento [CCIR, 1986-90c] informa sobre estudios efectuados en Alemania siguiendo la metodología que se describe en el § 3.1. Los resultados se muestran en la Figura 1. En estos ensayos participaron unos 55 evaluadores. Se efectuó también un análisis de los resultados de los diferentes grupos de edades (jóvenes/adultos) y regiones (Norte/Sur) en Alemania. Las diferencias surgidas eran relativamente insignificantes.

El Documento [CCIR, 1986-90d] informa sobre estudios efectuados en Francia siguiendo la metodología que se describe en el § 3.1. Los resultados se muestran en la Figura 1. Participaron unos 50 evaluadores que estaban en parte familiarizados con las escalas de calidad del CCIR y con los descriptores de escalas de degradación. La dispersión de los resultados en este caso era bastante inferior a la de los otros grupos lingüísticos.

En el § 3.1 se han descrito los resultados de los ensayos realizados en Estados Unidos y en Italia. Quedan también reflejados en la Figura 1. La tendencia más evidente puede observarse con tres de los cuatro gráficos en términos cualitativos (exceptuando Alemania) que figuran en el extremo inferior. Los términos se agrupan dejando únicamente un pequeño intervalo entre ellos. La escala de cinco notas y cuatro intervalos es en realidad una escala de cuatro notas y tres intervalos de separación distinta.

Los términos de la escala de degradaciones se distribuyen de forma bastante regular en los tres idiomas, debido quizás a que la sensación molesta se percibe como un fenómeno esencialmente.

3.4 Evaluación subjetiva de las relaciones de protección para las señales de radiodifusión de ondas decimétricas

En Estados Unidos se han realizado estudios para investigar cierto número de factores que inciden en la utilización de la banda de ondas decimétricas de la televisión por los servicios radioeléctricos móviles terrestres. Uno de esos estudios (Jones, 1986) trató de correlacionar la relación de protección deseada/no deseada con la calidad de la imagen. Los detalles del procedimiento empleado en las pruebas figuran en el documento de referencia. Se utilizaron escalas gráficas y de relación en condiciones de imágenes diferentes y similares.

Los resultados de estas pruebas indican que los telespectadores esperan más de la calidad de la imagen hoy día que en el pasado.

El Cuadro I muestra un ejemplo de los resultados de las pruebas, con las apreciaciones de los evaluadores para la condición de imagen correspondiente a la relación de 28 dB en las dos pruebas separadas, a saber, mediante valoración cuantitativa y mediante escalas gráficas en las que se puede ver una buena concordancia. La imagen tendría que mejorarse en un factor de tres para ser considerada como "aceptable" en la observación cotidiana.

CUADRO I

Señal deseada/señal no deseada = 28 dB
y desplazamiento de 10 kHz (525,24 MHz)

VALORACION CUANTITATIVA

el nivel "Aceptable" se fija arbitrariamente igual a 100.

	Media geométrica	Desviación típica geométrica
Expertos	= 35,5	2,9
No expertos	= 35,0	2,3
Todos los observadores	= 35,3	2,4

ESCALA GRAFICA

Expertos	= "Mediocre"
No expertos	= "No totalmente pasable"
Todos los observadores	= "No totalmente pasable"

3.5 Utilización de los descriptores de la escala de calidad del CCIR

Se han estudiado ampliamente los términos de la escala de calidad tradicionales de la Recomendación 500 del CCIR. Se han establecido en varios países e idiomas (Francia, Alemania, Estados Unidos e Italia) para determinar sus significados y la amplitud de los intervalos entre ellos. Las escalas gráficas resultantes se han utilizado satisfactoriamente para la evaluación subjetiva de la calidad de la imagen.

Se ha señalado que en un entorno de TVAD que utilice métodos de doble estímulo, los términos tradicionales no son aplicables. Se sugiere que los términos no pueden utilizarse de manera que se describa la calidad percibida de la imagen. Por ejemplo, como las imágenes de calidad inferior o degradadas no se consideran, ¿pertenecen a la escala los términos "mala" y "mediocre", incluso el de "aceptable"? en realidad, no se presentan a evaluación imágenes "malas" o "mediocres".

Se ha propuesto un experimento llevado a cabo en el que se establecían los términos de la escala tras las evaluaciones de la calidad de la imagen. [CCIR, 1986-90e] (escalas de términos post-producción). Se indicaba a los sujetos que establecieran en sus escalas cualquier término o todos los que sirviesen para describir las imágenes que acababan de ver. Observaban imágenes degradadas con calidad de estudio NTSC y de 1125 líneas.

Los sujetos situaron cada término en la escala sin excepción. Al preguntarles sobre las imágenes "malas" o "mediocres", dijeron que no había ninguna, aunque habían establecido los términos de la escala. Sucedió que, al comparar los grupos de imágenes de gran calidad con las imágenes de TVAD, las respuestas a las imágenes buenas tenían que desplazarse hacia abajo en la escala por las respuestas a las imágenes de TVAD, con lo que los significados de los términos perdieron relevancia en cuanto a la calidad de la imagen.

Uno de los objetivos de este estudio era comprobar si la situación de los términos de la escala después de evaluar la calidad de la imagen se adaptarían más a las imágenes contempladas. No fue así. Los términos se situaban en la escala exactamente igual que antes y no se relacionaban con la calidad de la imagen. Puede concluirse que se formuló una pregunta equivocada y que si se hubieran dado a los sujetos palabras para situar en la escala, las habrían ordenado por significados semánticos y no en relación con sus escalas de calidad de la imagen.

Se propone adoptar un enfoque binario: en primer lugar, preguntar al sujeto si ha visto imágenes que pueden describirse mediante cada palabra y que dé una respuesta de sí o no; en segundo lugar, clasificar todas las palabras que hayan recibido una respuesta positiva. Tal vez esto provoque una conexión más directa entre el significado de los términos y la calidad de la imagen. Otro enfoque sería pedir al sujeto que crease una palabra o que eligiese entre otras muchas, para describir una calidad fija y conocida de la imagen (desconocida para el sujeto).

4. Escala de categorías númericas

La escala de categorías numéricas se basa en la capacidad de los observadores para emitir juicios en categorías basadas en una escala lineal. Como las categorías no limitan su valor por los adjetivos, la escala puede utilizarse para gamas muy distintas.

El número de puntos en la escala que hay que elegir depende de las condiciones y de la gama de los atributos de percepción.

La experiencia ha demostrado que en algunos casos, son suficientes 5 puntos, mientras que en otros se necesita los puntos medios de una escala de 10.

Presenta algunas ventajas adoptar una escala a la que los observadores se habitúan en la vida cotidiana. Por ejemplo, en algunos países europeos una escala de 10 puntos representa una gama habitual en las escuelas.

Una escala de categorías númericas actúa más rápidamente y es fácil de automatizar (por ejemplo, 10 botones). Existen pruebas para la igualdad de los pasos a lo largo de la escala de evaluación (Edwards, 1957).

Si es necesario, los números pueden traducirse fácilmente a etapas de categorías de igual tamaño, y en las que el soporte lógico disponible utiliza los modelos de Thurstone (Torgersen, 1958).

Antes de empezar el experimento real, es aconsejable efectuar algunos ensayos en los que se muestre toda la gama, aunque sin especificación. Ello ayuda a los observadores a establecer sus escalas internas en el orden correcto. El número de ensayos por condición depende completamente del objetivo de la evaluación. No obstante, se aconsejan tres como mínimo para obtener un control estadístico.

5. Escalas multidimensionales

5.1 Introducción

El Documento [CCIR 1986-90f] informa sobre experimentos en que se pidió a los observadores que evaluaran la similitud entre imágenes con distintos grados de ruido, interferencia cocanal e interferencia de canal adyacente. A partir de los resultados se intentó establecer un espacio de percepción tridimensional. Esto sólo se pudo conseguir parcialmente, lo cual puede ser debido a efectos finales. Se están realizando nuevos análisis.

5.2 Métodos con escalas multidimensionales

Varios investigadores han utilizado métodos con escalas multidimensionales al considerar las apreciaciones de comparación de estímulos relativas a la televisión [Linde y otros, 1981; Goodman y Pearson, 1979]. Una prueba típica de escala multidimensional comienza con apreciaciones de comparación de estímulos (ya sea por categorías o no (véase la Recomendación 500)), sobre la similitud de los miembros de pares de condiciones. Seguidamente, se considera que las apreciaciones sobre el grado de similitud reflejan las "distancias" entre las condiciones en un espacio perceptual de n dimensiones y se aplica a las apreciaciones uno de varios procedimientos bien establecidos para determinar y designar las dimensiones de ese espacio [Shiffman y otros, 1981].

Tales métodos pueden contribuir a un enfoque en tres etapas para el estudio de la televisión. En primer lugar, la escala multidimensional puede utilizarse para establecer las dimensiones perceptuales en las que varían los factores de diseño y transmisión. En segundo lugar, pueden utilizarse las coordenadas de los niveles de factores en el espacio perceptual para definir relaciones entre parámetros objetivos y perceptuales. Y por último, las dimensiones perceptuales pueden relacionarse con apreciaciones de calidad o de satisfacción del espectador. Hasta ahora, no obstante, sólo se han utilizado métodos con escalas multidimensionales en un reducido número de estudios de la calidad de las imágenes de televisión. Se necesita proseguir las investigaciones para determinar el valor de estos métodos y del enfoque general. En [Lupker y Hearty, 1987)] se ofrece una descripción más completa de este enfoque. En Canadá se están realizando actualmente estudios de la utilidad del enfoque.

5.3 Método de variables múltiples

En el proyecto ESPRIT 925 se ha utilizado un método similar que se está estudiando en España [CCIR, 1986-90g]. Los experimentos tratan de identificar, agrupar e interpretar variables y factores subjetivos de interés que afectan a la calidad de la imagen. El diseño del experimento incluye un cuestionario en el que se hacen una serie de preguntas acerca de las secuencias de imagen observadas. Las cuestiones se refieren a aspectos tales como: cuáles son los atributos de mayor y menor preferencia; si las secuencias tienen calidad igual o distinta; si hay una degradación de la imagen y la forma en que puede corregirse; y cuáles son los rasgos, en orden de importancia, que conforman una imagen de gran calidad (para más detalles véase [CCIR, 1986-90 g]). Se vio que las variables más importantes o de mayor influencia eran:

- sonoridad de la señal audio

- imagen: bordes (difusos o netos)
  brillo y luminosidad
  nitidez
  movimiento (esp. horizontal)
  color
  ruido de imagen
  contraste
  parpadeo
  contenido global y local
  armonía de la composición
  nitidez de los aspectos faciales
  expresividad de los planos
  relación de claridad del motivo plano respecto al fondo
  continuidad de las secuencias
  posición de los sujetos

Las variables pueden entonces asignarse a factores globales tales como:

- contenido local y ruido

- contenido y planos: fondo

- calidad general (color, contraste, etc.)

- expresividad de los motivos

- nitidez de los motivos

- movimiento

o incluso además: contenido, ruido y color.

Continúa el trabajo para mejorar y ampliar el cuestionario y para establecer la validez del enfoque en las diversas evaluaciones de la calidad de la imagen de televisión.

5.4 Método de prueba ortogonal

En China, los estudios demuestran que la aplicación del método de prueba ortogonal para la evaluación subjetiva de la calidad de las imágenes de televisión, puede generalizarse, dentro de una determinada tolerancia de distorsión y mediante un pequeño número de experimentos, un grupo de combinaciones de distorsión que tienen las mismas características principales que una prueba completa. Además, la independencia de las distorsiones en cada combinación puede ser verificada, lo que proporciona un grupo razonable de combinaciones de distorsión para la evaluación subjetiva de la calidad de las imágenes afectadas por cinco distorsiones simultáneas [CCIR, 1986-90h].

6. Escala gráfica con presentación de triple estímulo

Los resultados experimentales muestran que la evaluación de la imagen con escalas de categorías puede no dar un orden de las escalas. Los métodos de escala gráfica pueden constituir una alternativa. El Documento [CCIR 1986-90i] describe un enfoque de escala gráfica que permite controlar el tipo de escala resultante (escala ordinal o de intervalos). En un experimento de presentación de triple estímulo, se indicó a los observadores que señalasen gráficamente en un monitor central la situación de la calidad de la imagen, en relación con la calidad de las imágenes de dos monitores situados a cada lado. Las imágenes se degradaron con niveles distintos de ruido. Se visualizaron todas las combinaciones posibles de un conjunto de niveles de ruido. Este método se ha utilizado para demostrar que las evaluaciones según escalas subjetivas pueden no dar una escala de intervalos.

7. Procedimiento de comparación por parejas

7.1 Descripción general

Este método puede dar únicamente un orden de categorías de imágenes o secuencias conforme a su calidad subjetiva, que se obtiene a partir de las evaluaciones de todas las parejas posibles. El método presenta la ventaja de que los datos facilitan un examen del carácter transitivo de los juicios que emiten cada uno de los sujetos y de la concordancia de los criterios utilizados por ellos. No siempre es evidente que se satisfacen estas condiciones, especialmente cuando intervienen degradaciones de imágenes complejas. Si los resultados del examen son negativos, debe considerarse un método de evaluación multidimensional (escalas multidimensionales o análisis de factores).

La fiabilidad del método depende del número de imágenes o de secuencias (que no debe ser inferior a seis) y del número de sujetos.

7.2 Procedimiento de prueba

Un número N de sujetos debe establecer un orden de categorías de n imágenes o secuencias. A los sujetos se le presentan todas las parejas posibles en orden aleatorio. Deciden para cada pareja cuál es la imagen o secuencia mejor. El número total de parejas es:

$$z = n\,(n-1)/2$$

Los resultados de la prueba se combinan en una matriz de dominancia individual (n x n) para cada sujeto. Un "uno" en la columna t y en la línea s indica que la imagen o secuencia t es mejor que la s, y un "cero" indica lo contrario.

7.3 Examen de la transitividad individual (método Zeta)

Debe verificarse la transitividad de los resultados de todos los sujetos. Un sujeto habrá efectuado un "ensayo intransitivo" si decidió que la imagen A era mejor que la B, la B mejor que la C, pero que la C era mejor que la A. Un orden de categorías solo se puede obtener a partir de los resultados de la prueba si los sujetos opinan de un manera sistemáticamente transitiva.

El número de ensayos intransitivos en una matriz de dominancia es:

$$d = (n(n - 1)\ (2n - 1)/12 - 1/2) \sum_i D_i^2$$

en la que D_i es la suma de la columna i-ésima.

El máximo de d es:

$d_{máx} = n\ (n^2 - 4)/24$ si n es par

$= n\ (n^2 - 1)/24$ si n es impar

ζ es una medida de la transitividad:

$\zeta = 1 - d/d_{máx}$

Si ζ es igual a 1, hay transitividad absoluta. Si ζ es igual a 0, los juicios son completamente intransitivos.

Los resultados transitivos pueden ser también aleatorios. Verificando la probabilidad de que el valor calculado de los ensayos intransitivos (d) con la condición de que la transitividad sea exclusivamente aleatoria, puede adoptarse una decisión en cuanto a si las evaluaciones del sujeto pueden considerarse sistemáticamente transitivas o no. En primer lugar, se fija un límite α para la probabilidad, que es suficientemente pequeño, por ejemplo, 0,05 (5%). A continuación se calcula la fórmula siguiente:

$$x = \frac{8}{n - 4}\left(1/4 \binom{n}{d} - d + 1/2\right) + DF$$

$DF = n\ (n - 1)\ (n - 2)\ /\ (n - 4)^2$ (grados de libertad)

Con la condición de que $n > 6$, x (d) es una función con distribución χ^2. La χ^2 es una función de distribución de probabilidad muy conocida (Cuadro II). Para un valor fijo de α y el valor calculado de DF, el valor de χ^2 viene dado en el cuadro II. Si esta valor es inferior al valor calculado de x, se supone que la transitividad es sistemática.

CUADRO II

Distribución χ^2

Grados de Libertad	α = ,05	α = ,01	α = ,001
1	3,84	6,63	10,83
2	5,99	9,21	13,82
3	7,81	11,34	16,27
4	9,49	13,28	18,47
5	11,07	15,09	20,52
6	12,59	16,81	22,46
7	14,07	18,48	24,32
8	15,51	20,09	26,13
9	16,92	21,67	27,88
10	18,31	23,21	29,59
11	19,68	24,73	31,26
12	21,03	26,22	32,91
13	22,36	27,69	34,53
14	23,68	29,14	36,12
15	25,00	30,58	37,70
16	26,30	32,00	39,25
17	27,59	33,41	40,79
18	28,87	34,81	42,31
19	30,14	36,19	43,82
20	31,41	37,57	45,32
21	32,67	38,93	46,80
22	32,92	40,29	48,27
23	35,17	41,64	49,73
24	36,42	42,98	51,18
25	37,65	44,31	52,62
26	38,89	45,64	54,05
27	40,11	46,96	55,48
28	41,34	48,28	56,89
29	42,56	49,59	58,30
30	43,77	50,89	59,70
40	55,8	63,7	73,4
50	67,5	76,2	86,7
60	79,1	88,4	99,6
70	90,5	100,4	112,3
80	101,9	112,3	124,8
90	113,1	124,1	137,2
100	124,3	135,8	149,4

ARCHIVES U.I.T. GENÈVE

7.4 Examen de la concordancia de los sujetos

El cálculo de un orden de categorías común es únicamente razonable si los sujetos utilizan los mismos criterios en sus opiniones (es decir, si su concordancia es sistemática). Para examinar esto, se combinan los resultados de las pruebas en una denominada matriz de agregados. A este efecto, se enumeran todas las parejas de imágenes o secuencias de prueba. El orden es arbitrario. Estos números corresponden a los números de las líneas de la matriz. Los números de las columnas corresponden a los números (arbitrarios) de los sujetos. El elemento X_{ij} de esta matriz es igual a 1 (0), si el sujeto j opina que la primera imagen de la pareja i es mejor (peor) que la segunda.

La concordancia entre los sujetos puede ser sistemática o aleatoria. Se supone que es sistemática si la probabilidad de concordancia real es suficientemente pequeña, a condición de que la concordancia sea únicamente aleatoria. En primer lugar, se fija un límite α para la probabilidad, que es suficientemente pequeño, por ejemplo, $\alpha = 0{,}05$ (5%). A continuación se calcula el valor siguiente:

$$Q = \frac{\binom{n}{2}\left[\binom{n}{2} - 1\right] \sum_i (L_i - \bar{L})^2}{\binom{n}{2} \sum_j G_j - \sum_j G_j^2}$$

n: número de imágenes o secuencias
N: número de sujetos
L_i: suma de la línea i-ésima de la matriz de agregados

$$\bar{L} = \Sigma L_i / \binom{n}{2}$$

ΣG_j: suma de la columna j-ésima de la matriz de agregados.

Además, se calcula el número de grados de libertad:

$$DF = \binom{n}{2} - 1$$

Para el valor fijo de α y el valor calculado de DF, el valor correspondiente de χ^2 viene dado en el Cuadro II. Si Q es mayor que χ^2, se supone que la concordancia es sistemática.

7.5 Cálculo del orden de categorías

De los resultados de prueba puede obtenerse un orden de categorías a condición de que la transitividad de todos los sujetos y la concordancia entre ellos sean sistemáticas.

Se calcula una matriz de dominancia global sumando las matrices individuales. Un elemento X_{ij} de esta matriz es igual a la frecuencia de la opinión de que la imagen j es mejor que la imagen i. A continuación, se calculan las sumas de columnas D_i de esta matriz. D_i es la frecuencia de la opinión de que la imagen i es mejor que cualquier otra imagen. El orden de categorías de imágenes o secuencias viene dado por el orden de estas sumas de columnas.

8. Método de evaluación de umbrales de visibilidad

Para determinadas mediciones y sobre todo para obtener el mayor grado de precisión cuando se establece una correspondencia entre mediciones objetivas y evaluación de su influencia en la calidad visual, resulta interesante medir los umbrales de visibilidad de "factores de degradación". Incluso si el método del doble estímulo con escala de degradación puede aportar información a este respecto, parece conveniente preconizar el empleo de un método más sencillo y mejor llamado "método del doble estímulo con elección forzada". La eficacia de este método, utilizado a menudo en el marco de estudios psicofísicos, se ha verificado en el contexto de los estudios realizados en el CCIR. Los resultados se muestran en [1986-90j].

El procedimiento utilizado para material natural o sintético se basa en la comparación de una secuencia degradada con la referencia correspondiente. En cada par de secuencias que constituyen una presentación, la posición de la referencia es aleatoria. El cometido de los observadores consiste únicamente en indicar cuál de las dos secuencias de la presentación está degradada. Se dice que la elección es forzada porque los observadores deben siempre dar una respuesta, incluso cuando tengan dudas.

Los niveles de degradación presentados deben abarcar una gama suficientemente amplia por encima y por debajo del umbral de visibilidad estimado.

El tratamiento de las evaluaciones sigue el protocolo siguiente: puesto que la probabilidad de respuestas acertadas varía entre el 50% (correspondiente a la no percepción de degradación, es decir, respuestas aleatorias) y el 100% (correspondiente a la degradación observada siempre), suele estimar el umbral de visibilidad en el 75% para cada observador. Una vez estimado el umbral de cada observador, se calcula el umbral medio y su correspondiente intervalo de confianza. Puede elegirse otro porcentaje para medir un umbral menos riguroso.

Se ha verificado la estabilidad del método, así como su aptitud para facilitar un verdadero umbral de visibilidad en comparación con el método del doble estímulo que utiliza una escala de degradación. Los resultados están claramente a favor del método del doble estímulo con elección forzada cualquiera que sea la gama de degradaciones elegidas, pero es preferible presentar niveles de degradación correctamente repartidos en torno al umbral estimado.

9. Etapas para la incorporación de un nuevo método en la Recomendación 500

Con los datos disponibles actualmente, si hay que eliminar o ampliar los métodos actuales de la Recomendación 500, el candidato más probable es un método de valoración cuantitativa según las líneas marcadas en el punto 2.1.

Los resultados de un ensayo muestran que el método es menos sensible al contexto que un método de escala de calidad con un solo estímulo, y un segundo ensayo indica que la correlación entre laboratorios sería buena siempre que las escalas de calidad de la imagen de referencia sean las mismas.

Estos resultados alentadores han de confirmarse por una serie de laboratorios con diferencias lingüísticas.

También es importante ofrecer pautas en cuanto a cómo debe interpretar los resultados la comunidad de radiodifusión. En dicha comunidad se suele trabajar con escalas de cinco notas y se requiere una explicación de la relación entre los dos entornos.

Deben continuar también los estudios sobre métodos de escalas de categorías numéricas, escalas multidimensionales y escalas gráficas (que se describen en el punto 4 de la Recomendación 500 con miras a establecer las ventajas que puedan presentar respecto a otros métodos alternativos.

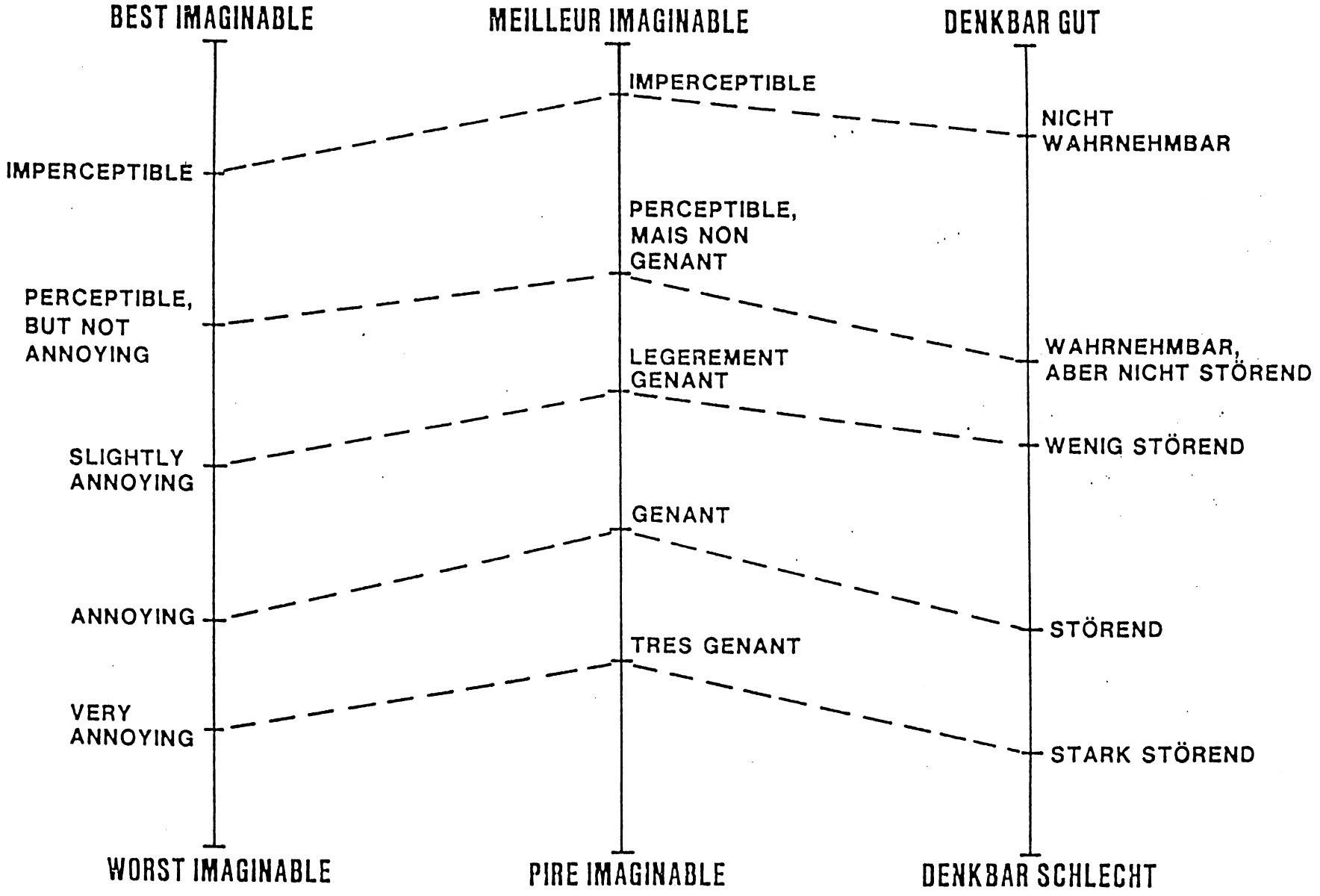

FIGURA 1a

Escalas gráficas de términos de calidad

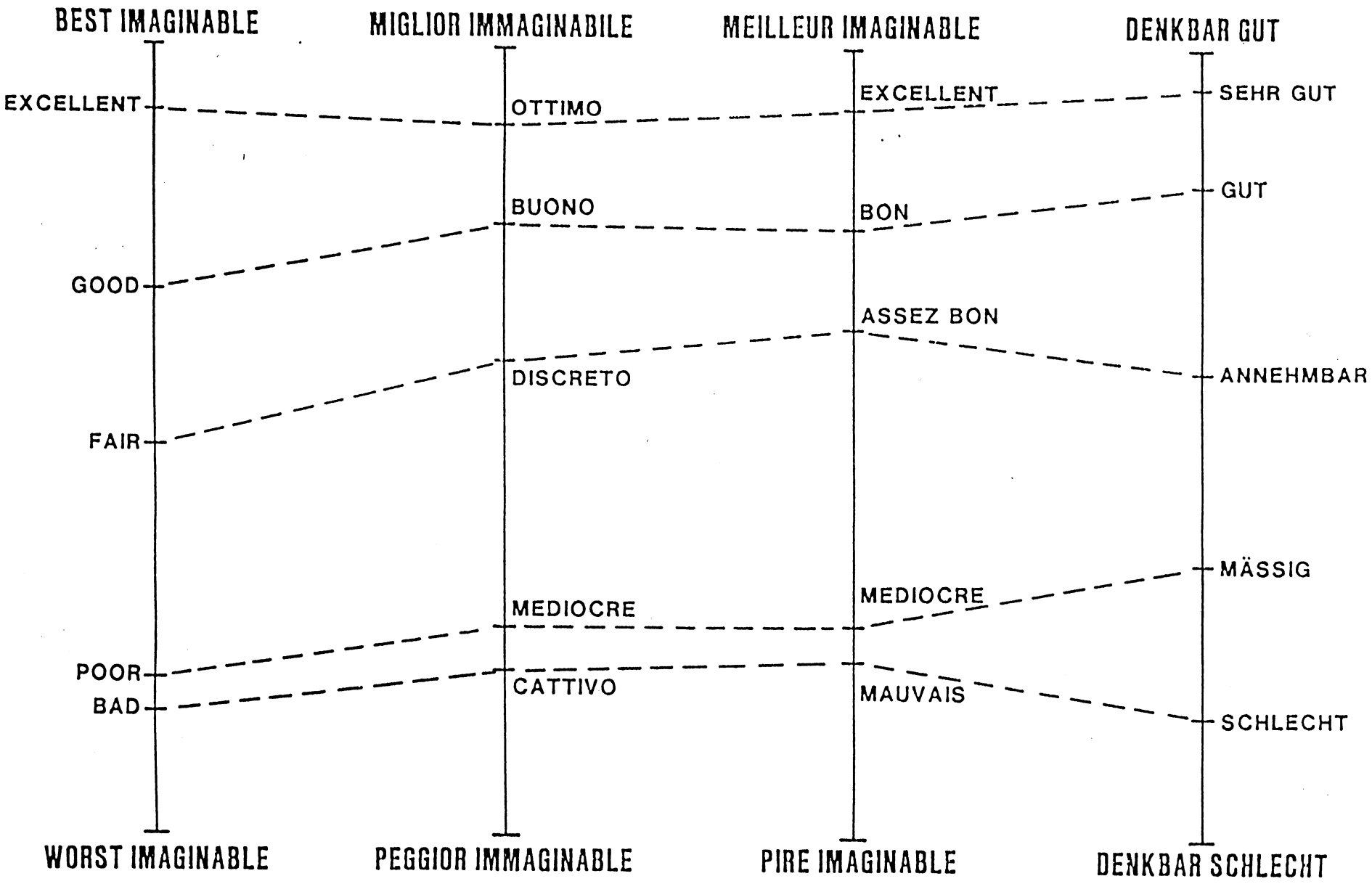

FIGURA 1b

Escalas gráficas de términos de degradación

REFERENCIAS BIBLIOGRÁFICAS

GOODMAN, J.S. y PEARSON, D.E. [1979] - Multidimensional scaling of multiply-impaired television pictures. IEEE Trans. Systems, Man. Cybernetics, 9, 353-356.

JONES, B.L. y McMANUS, P.R. [1986] - Graphic scaling of qualitative terms. SMPTE J., Vol. 95, 11-86, 1166-1171.

JONES, B.L. [11, julio, 1986] - Subjective assessment of protection ratios for UHF broadcast signals. Study field as NAB comments to the FCC, General Docket, No. 85-172.

LINDE, L., MARMOLIN, H. y NYBERG, S. [1981] - Visual effects of sampling in digital picture processing - A pilot experiment. IEEE Trans. Systems, Man. Cybernetics, 11, 201-207.

LUPKER, S. y HEARTY, P. [1987] - Evaluating the effects of multiple sources of impairment in TV signals. Proc. 3rd International Colloquium on Advanced Television Systems: HDTV'87, Ottawa, Canada.

SCHIFFMAN, S.S., REYNOLDS, M.L. y YOUNG, F.W. [1981] - Introduction to multidimensional scaling. New York, Academic Press.

EDWARDS, A.L. [1957] - Techniques of attitude scale construction. NY Appleton Century Crofts Inc.

TORGERSON, W.S. [1958] - Theory and methods of scaling. NY John Wiley & Sons.

Documentos del CCIR

[1986-90]: a. 11/379 (USA); b. GIT 11/4-123 (Francia); c. GIT 11/4-137 (República Federal de Alemania); d. GIT 11/4-147 (Francia); e. GIT 11/4-160 (Presidente, GIT 11/4); f. GIT 11/4-145 (Canadá); g. 11/158 (España); h. 11/144 (República Popular de China); i. GIT 11/4-141 (República Federal de Alemania); j. 11/463 (Francia).

ANEXO I

Condiciones para la evaluación durante la transmisión del programa

Referencias	OIRT [CCIR, 1966-69 a]		Canadá [CCIR, 1966-69 b]
Observadores			
Categoría	Especialistas		Especialistas
Número	1 ó 2		1 ó 2
Escala de apreciación			
Tipo	Degradación	Calidad	Degradación
Número de notas	6 (Nota 1)	6 (Nota 2)	5 (Nota 3)
Imagen			
Tipo	Programas de televisión		Programas de televisión
Condiciones de observación			
Relación entre la distancia de observación y la altura de la imagen	4-6		4-6
Ángulo de visión con relación a una línea perpendicular al receptor de control			<30°
Luminancia en la pantalla para el blanco de referencia (cd/m²)			70 ± 7
Cromaticidad de la pantalla para el blanco de referencia			Iluminante *D*
Luminancia de la pantalla del tubo inactivo	Adaptado a la iluminación ambiente		Lo más débil posible
Luminancia del «recuadro luminoso» (cd/m²)			10,5 ± 3,5 (Nota 14)
Cromaticidad del «recuadro luminoso»			Iluminante *D*

Nota 1.- *Escala de degradación* de 6 grados

1 imperceptible
2 apenas perceptible
3 claramente perceptible, pero no molesta
4 ligeramente molesta
5 claramente molesta
6 inutilizable

Nota 2.- *Escala de calidad* de 6 grados

1 excelente
2 buena
3 aceptable
4 mediocre
5 mala
6 muy mala

Nota 3.- *Escalas de degradación* de 5 grados

1 imperceptible (implícita)
2 perceptible
3 apreciable
4 molesta
5 inutilizable

REFERENCIAS BIBLIOGRÁFICAS

Documentos del CCIR

[1966-1969]: a. XI/46 (OIRT); b. XI/146 (Canadá).

ANEXO II

Información adicional sobre procedimientos de prueba

En este anexo se describen y analizan varios procedimientos para organizar pruebas de evaluación subjetiva. En la Recomendación 500 figura abundante información sobre la evaluación subjetiva y sobre los procedimientos de prueba más corrientemente empleados. En el punto 1 del presente anexo se describen procedimientos suplementarios y nuevos. En el punto 2 se describe la técnica de la muestra virtual. En el punto 3 se da información sobre los resultados obtenidos por algunos de los procedimientos en situaciones comparables. El punto 4 contiene algunas pruebas adicionales. En el punto 5 se examinan diferentes argumentos relacionados con la elección de un método adecuado para diversas situaciones.

1. Descripción de los procedimientos

En este punto se presentan descripciones de procedimientos adicionales a los descritos en la Recomendación 500 entre ellos algunos destinados a resolver nuevos problemas. Dichas descripciones abarcan puntos tales como las escalas de apreciación, la determinación de la secuencia de presentación, la definición de una condición degradada, etc. La validez de estos puntos no está limitada necesariamente al procedimiento particular descrito.

1.1 *Procedimiento que utiliza la escala de calidad con enclavamiento directo**

Para estudiar la función de una imagen de referencia se ha diseñado un nuevo método [Kretz y Sallio, 1981]: En este método se utiliza la escala de calidad de cinco notas de la Recomendación 500 y el procedimiento es idéntico al del método de la UER (véase más adelante el punto 1.2): se proyecta una imagen de referencia antes de cada imagen que se va a evaluar y se pide a los observadores que califiquen las imágenes con relación a la imagen de referencia, a la que se supone corresponde la nota 5 («excelente»). Con este procedimiento, se obtiene un enclavamiento directo de las notas en lo alto de la escala con la escala de apreciación de la calidad.

1.2 *Procedimiento que utiliza secuencias de imágenes animadas*

La Recomendación 500 especifica las condiciones en que ha de efectuarse la evaluación subjetiva de la calidad de las imágenes fijas y de las animadas. Sin embargo, hay pocos estudios que traten de alguna de las características básicas de la medición de la calidad subjetiva de las secuencias de imágenes animadas. Esas características comprenden:

- los elementos en los que el observador basó su decisión suelen ser transitorios;
- al observador le puede resultar difícil analizar y evaluar todos los elementos de la secuencia en una sola visualización;
- la percepción de ciertas degradaciones puede ser diferente según se trate de imágenes fijas o de imágenes animadas.

En un primer intento de definir un procedimiento adecuado, un estudio [CCIR, 1982-86a] en el que se utiliza la escala de degradación con una imagen de referencia (método de la UER) ha examinado diversos modos de presentación de secuencias de imágenes animadas. La degradación utilizada consistió en generaciones múltiples de grabaciones magnetoscópicas.

Pueden sacarse las siguientes conclusiones preliminares sobre la base de las puntuaciones medias y de las desviaciones típicas. En primer lugar, las secuencias de más de 10 s de duración parecen ser demasiado largas y, en segundo lugar, la repetición de las secuencias con el objeto de dar a los observadores una mayor oportunidad de analizar y evaluar las escenas no parece mejorar la calidad de las evaluaciones. La conclusión de este estudio preliminar parece ser que es preferible una sola presentación de unos 8 – 10 s. Además, los resultados indican que las evaluaciones de las mismas degradaciones en secuencias de imágenes fijas y animadas pueden diferir de manera significativa.

Las precedentes conclusiones son de carácter preliminar y se basan en resultados obtenidos con un solo tipo de degradación. Con degradaciones dependientes del movimiento, esos resultados podrían diferir significativamente. Urge efectuar estudios ulteriores sobre este importante aspecto.

---

* En todos los procedimientos hay implícita una cierta forma de enclavamiento en el sentido de que es necesario normalizar el proceso de encaminamiento. La expresión «enclavamiento directo» se refiere aquí al enclavamiento *explícito*. El «enclavamiento indirecto» corresponde a la normalización de los fenómenos de adaptación [Corbett, 1970] por medio de la gama de degradaciones en una sesión de prueba.

2. La técnica de la muestra virtual

2.1 Las principales fuentes de errores en las pruebas subjetivas son esencialmente dos:

- los errores aleatorios (es decir, estocásticos); y
- los errores sistemáticos.

Si se han normalizado las condiciones objetivas de la prueba, la naturaleza de los errores sólo guarda relación con los parámetros utilizados en el diseño de las pruebas (número de observadores, imágenes, procedimiento utilizado, etc.).

Los errores estocásticos se reconocen con gran facilidad. En el caso de una relación entre la magnitud de distorsión y la calidad de la imagen, conducen a cierto grado de dispersión de las notas medias experimentales en torno a la curva ajustada (es decir, obtenida por el método de los mínimos cuadráticos). La normalización de los procedimientos de prueba suele tender a reducir los errores estocásticos; pero debido a su naturaleza pueden reducirse más aún por promediación estadística (la utilización de una función matemática ajustada es un tipo de promediación).

Los errores sistemáticos son difíciles de reconocer puesto que casi nunca guardan relación con los errores aleatorios. Generalmente actúan desplazando la curva anteriormente mencionada (sesgo) y/o afectando a su pendiente. Una vez introducido un cierto grado de errores sistemáticos en resultados experimentales, no pueden ser promediados por métodos estadísticos.

2.2 La «muestra virtual» está constituida por un número relativamente elevado de observadores (50 o más) y también por un gran número de imágenes (40 o más). Se llama «muestra virtual» porque no se utiliza en su totalidad para la prueba de evaluación real sino que se utiliza como una población de la que se extraen repetidas muestras más reducidas y de un tamaño manejable.

Tomando un ejemplo en el que se da una relación entre determinadas magnitudes de distorsión y calidad de la imagen y teniendo en cuenta que la finalidad de la técnica de la muestra virtual es utilizar diferentes muestras de observadores e imágenes para diferentes condiciones de prueba, la planificación completa del experimento, siguiendo la técnica de la muestra virtual, sería la siguiente:

- se selecciona un número de condiciones de prueba (por ejemplo, de 8 a 10 valores de la magnitud de distorsión);
- se forma un número de grupos, integrado cada uno por 2 ó 3 condiciones no contiguas de prueba;
- para cada condición de prueba se escoge de la serie completa una muestra aleatoria de 5 a 6 imágenes de prueba (muestra virtual), de forma que cada grupo de condiciones de prueba tenga 2 ó 3 series de imágenes de prueba. Para cada grupo se selecciona una muestra de 8 a 10 observadores (muestra virtual). De esta manera tenemos diferentes imágenes de prueba para diferentes condiciones de prueba en el mismo grupo;
- para cada grupo de condiciones de prueba se organiza una sesión siguiendo el procedimiento descrito en el punto 8.4 del apéndice a la Recomendación 500-3, Recomendaciones e Informes del CCIR, Vol.XI-1, Dubrovnik, 1986;
- se calculan y ajustan después las notas medias por el método de los mínimos cuadráticos utilizando la función adecuada (por ejemplo, la función logística);
- habrá que realizar alguna prueba estadística de la bondad del ajuste y el resultado final del experimento será la curva ajustada que se ha obtenido.

Si se requiere mayor exactitud para la curva ajustada, podrá repetirse de nuevo el experimento y la nota media correspondiente se promediará con las anteriores, antes de proceder al nuevo ajuste por el método de los mínimos cuadráticos.

Hay un acuerdo general respecto a que la técnica es adecuada para degradaciones complejas, al menos en relación con los observadores. Sin embargo, algunas administraciones consideran que puede no ser económica en el caso de una sola degradación.

3. Resultados de experimentos directamente comparables

Se describen a continuación resultados obtenidos con diferentes procedimientos de evaluación aplicados al mismo material experimental (imágenes y degradaciones). Naturalmente, cuando se utiliza un método en una situación específica, se deben tener en cuenta numerosos parámetros y las conclusiones de tales experimentos, descritos en este punto, sólo tienen en cuenta algunos de esos factores. Ello se discute con mayor detalle en el punto 5.

3.1 *Comparación de los resultados obtenidos por el método de un solo estímulo y por el método de la UER*

Los dos procedimientos descritos en los puntos 8.1 y 8.2 del apéndice a la Recomendación 500-3, Recomendaciones e Informes del CCIR, Vol.XI-1, Dubrovnik, 1986: ——————————— han sido sometidos a una serie de experimentos comparativos [CCIR, 1978-82a, b, c; Sallio y Kretz, 1982]. Se utilizaron diferentes tipos de degradación: filtrado analógico, ruido aditivo (en dos contextos diferentes), ruido de transferencia por modulación cruzada, borrosidad en los bordes (en dos contextos diferentes), acumulación de ruido aditivo y borrosidad en los bordes, errores de transmisión a 34 Mbit/s (de dos escalas de degradación diferentes). Participaron en las pruebas veinte grupos independientes de diez observadores no expertos (un grupo para cada tipo de degradación y procedimiento). Hubo un total de 46 sesiones para cada método. La comparación de los resultados obtenidos se analizó en términos de notas medias y desviación típica, separadamente, para cada distancia de observación. Se llegó a las siguientes conclusiones:

- con los dos métodos se obtienen curvas de relación objetiva-subjetiva de formas muy similares, pero hay una variación entre las características obtenidas por cada método. Las curvas de evaluación de la calidad tienden a estar más bien por debajo de las curvas de degradación (método de la UER). En la mitad de la escala (nota 3) las desviaciones típicas de las opiniones son máximas y muy próximas para ambos procedimientos (en 6*H*, 0,84 para el método de un solo estímulo y 0,79 para el método de la UER);
- el procedimiento de escala de degradación que utiliza una imagen de referencia produce una nota media para la imagen de referencia muy próxima a la nota más alta (4,88 como promedio), mostrando un buen enclavamiento para las evaluaciones en la parte superior de la escala;
- el procedimiento de la escala de calidad produce una nota media para las imágenes de referencia que varía de casi 0 a 1 por debajo de la nota más alta (4,56 como promedio). Esto parece deberse a la naturaleza absoluta de la evaluación;
- los dos métodos probados son sensibles al contexto y a la gama de las degradaciones presentadas en una sesión. Estos dos fenómenos subjetivos tienen efecto idéntico en los dos métodos. Por tanto sería importante, al presentar los resultados, especificar las condiciones precisas (gama de degradaciones presentadas en cada sesión, contexto en una sesión), lo que permitiría interpretar mejor los resultados;
- en la gama entre «imperceptible» y «perceptible pero no molesto» (alrededor de la nota 4,5), el procedimiento que utiliza la escala de degradación y una referencia, produce desviaciones típicas de las notas 1,4 veces inferiores a las obtenidas con el procedimiento de la escala de calidad; de este modo, el primer procedimiento parece dar una mejor precisión y podría reducirse a la mitad el número de notas hacia la parte superior de la escala.

Estos resultados muestran que puede ser posible obtener una transformación de grados medios de calidad en grados medios de degradación, obtenidos con el procedimiento de la UER variando los resultados experimentales por la magnitud asociada con la degradación residual (notas medias para imágenes degradadas). Se sugiere la transformación de las notas medias de degradación en notas medias de calidad variando los valores experimentales en media nota, aunque este valor no es completamente estable. ——————————— Debido a la naturaleza limitada de la escala, esta transformación no puede aplicarse en la parte inferior de la misma.

3.2 *Comparación de los resultados obtenidos con otros métodos*

A fin de estudiar más detalladamente la función de las imágenes de referencia, el enclavamiento y la escala de apreciación, se han probado varios métodos sobre el mismo tipo de degradación [CCIR, 1978-82b]. Este estudio consistió en la comparación de los resultados obtenidos con dos métodos mencionados anteriormente (véase el punto 3.1) y los resultados obtenidos con algunos otros métodos. Se estudiaron los siguientes aspectos:

- la utilización de una escala de degradación de cinco notas y de una escala de degradación continua, de la escala de apreciación de cinco notas, y de una escala de calidad continua, todas con una imagen de referencia para enclavamiento directo (de hecho, el procedimiento de la UER pero utilizando escalas diferentes);
- la utilización de una escala de calidad continua con un procedimiento de doble estímulo próximo al descrito en el punto 8.3 del apéndice a la Recomendación 500-3, Recomendaciones e Informes del CCIR, Vol.XI-1, Dubrovnik, 1986;

Se llegó a las siguientes conclusiones:

- es posible un buen enclavamiento en la parte superior de la escala con una presentación del tipo UER, pero utilizando una escala de apreciación de cinco notas e informando a los observadores que la imagen de referencia debe corresponder a la nota «excelente»;

- la comparación de los resultados obtenidos mediante métodos que difieren solamente en la utilización de escalas de apreciación discreta o continua demuestra que ni la escala de calidad continua ni la escala de degradación continua proporciona más información que la escala de cinco notas recomendada (valores medios y desviaciones típicas comparables);
- con el método del doble estímulo que no proporciona enclavamiento directo, la nota media para referencia no está próxima a la parte superior de la gama; las desviaciones típicas obtenidas mediante este método no son significativamente más bajas que las obtenidas mediante el método recomendado por la UER;
- la aplicación de escalas continuas plantea algunos problemas: para algunos observadores (no expertos) es difícil utilizarla y se hace más complicado analizar y presentar los resultados de los experimentos.

3.3 *Comparación de los resultados obtenidos utilizando la escala de calidad de cinco y seis notas utilizando la escala de degradación de seis notas*

En [Allnatt y Corbett, 1974] se ha informado sobre la comparación de los resultados obtenidos utilizando la escala de calidad de cinco notas y la escala de degradación de seis notas, y recientemente se ha examinado de nuevo esta comparación con referencia particular a la calidad cerca del umbral de visibilidad [Allnatt, 1980]. El procedimiento fue el mismo, salvo las escalas utilizadas. La escala de degradación de 6 notas difiere de la actual Recomendación 500. Se consideraron dos tipos de degradación: una con la televisión monocroma de 625 líneas utilizando un eco no distorsionado de 2 μs, y la otra con fotografías opacas degradadas por pérdida de definición. Los resultados se analizaron solamente en términos de nota media de opinión. La conclusión principal obtenida de este estudio es que la escala de degradación no ofrece ventajas respecto a la sensibilidad con degradaciones por debajo del umbral de visibilidad. Sin embargo, debe señalarse que el experimento no representa los resultados que se habrían obtenido comparando el método de un solo estímulo y el método de la UER.

4. Algunos otros hechos experimentales

Se dispone de hechos experimentales de otras fuentes, algunas de las cuales difieren de las del punto 3. En el caso del método de apreciación de la calidad con un solo estímulo, existe una cantidad considerable de resultados referentes a todas sus propiedades importantes.

Los resultados sustentan la aplicabilidad de la transformación de degradaciones (véanse, por ejemplo [Macdiarmid y Allnatt, 1978]; también el anexo II) como una ley de adición de degradaciones subjetivas. Esta ley se utiliza para ajustar el efecto de degradación residual en el análisis y funciona de manera diferente a la transformación que utiliza la variación de las notas medias expuesta en el punto 3.1.

En relación con el método de apreciación de la calidad de doble estímulo, se ha hallado que las desviaciones típicas de las diferencias entre pares de notas (en la misma presentación) son más bajas que las de las notas individuales, cuando las degradaciones son pequeñas. Se necesita una transformación para dar una desviación típica equivalente de la escala de cinco notas en degradación cero y se obtuvieron valores de aproximadamente 0,13 con una gama bastante amplia de degradaciones utilizando ruido aleatorio [White y Allnatt, 1980], y 0,35 en la evaluación de un códec horizontal de alta calidad [CCIR, 1978-82d]. En otros experimentos con televisión digital [IBA, 1981], el valor obtenido fue de 0,22. En otros experimentos similares [Kretz y Sallio, 1981] los valores comparables fueron 0,25 a $4H$, 0,45 a $6H$ para un nivel nulo de degradación.

La comparación de mediciones efectuadas con un códec digital utilizando el método de doble estímulo, y las obtenidas por el método de un solo estímulo [CCIR, 1978-82d], con las degradaciones a evaluar sumergidas en un gran número de nuevas degradaciones, confirma las conclusiones de White y Allnatt [1980], según las cuales la adaptación debida a los efectos de enclavamiento indirectos se reduce considerablemente utilizando el método de doble estímulo.

Los resultados obtenidos utilizando una escala de degradación muestran que, cuando se desea medir la visibilidad de las degradaciones aisladas, no conviene presentarlas en una misma secuencia [CCIR, 1982-86b]. Esta manera de proceder puede introducir un sesgo, pues los observadores tienen tendencia a comparar el efecto de las diferentes degradaciones. En el caso de la medición de visibilidad de degradaciones aisladas parece, pues, preferible juzgar un solo tipo de degradación por secuencia de presentación.

5. Análisis

La finalidad de especificar un procedimiento en detalle es minimizar la variación aleatoria de los resultados que no se debe a diferencias sistemáticas entre grupos diferentes de observadores cuando, por ejemplo se comparan o combinan los resultados de pruebas independientes. Esta es la finalidad común de todos los procedimientos descritos, pero no obstante existen otros factores que parecen dar solidez a determinados procedimientos. Estos factores se relacionan con el grado de discriminación de los resultados que vale la pena obtener, y el grado de complejidad y laboriosidad del procedimiento. Mientras más complejo y elaborado es el procedimiento, consumirá más tiempo y será más costoso. La precisión necesaria y la ganancia en términos de los resultados logrados son aspectos básicos para el análisis de la elección del procedimiento.

Los procedimientos descritos en el punto 8 del apéndice de la Recomendación 500-3, Recomendaciones e Informes del CCIR, Vol.XI-1, Dubrovnik, 1986; ——— reflejan una combinación de la escala de calidad o de degradación con la utilización regular de una imagen de referencia, que tiene significado o no como tal, o por un enclavamiento indirecto por gamas de degradación. El método descrito en el punto 8.1 es el más sencillo en cuanto a la organización; el análisis de los resultados de los métodos descritos en los puntos 8.1, y 8.2 ——— y en el punto 1.1 de este anexo tiene aproximadamente el mismo grado de complejidad; la organización de los métodos expuestos en los puntos 8.2 y 8.3 del apéndice ——— y en el § 1.1 de este anexo es aproximadamente igual, pero el análisis de los resultados para el método del § 8.3 toma más tiempo. El método descrito en el § 8.4 por lo general requiere más sesiones que los otros procedimientos y está diseñado para reducir los errores sistemáticos. En el punto 3 figura una comparación de los resultados obtenidos mediante algunos de los métodos para determinadas degradaciones, y en el punto 4 de este anexo se exponen otros hechos experimentales.

Es evidente que cada método parece tener algunos fundamentos sólidos, por lo que la elección entre ellos no es sencilla. Es imposible justificar plenamente los argumentos que se presentan en un Informe de esta clase, pero esencialmente, los principales argumentos que han influido sobre los experimentadores en el campo son los que se exponen a continuación.

La elección del procedimiento está relacionada con la elección de la escala de apreciación, bien sea continua o discreta, y con la manera en que debe pedirse a los observadores que hagan uso correcto de la escala.

Respecto al valor relativo de la escala de calidad y de la escala de degradación, algunos consideran que el concepto «calidad» está más estrechamente relacionado con los intereses de los observadores, y es además beneficioso, porque si una «degradación» mejora realmente la imagen, los resultados lo reflejan. Por otra parte, otros consideran que es más fácil interpretar la escala de degradación y que tiene la ventaja de permitir la medición de un umbral de percepción (entre las notas 4 y 5 de la escala de degradación). Parecería que existe un equilibrio de preferencias en la opinión de los observadores con respecto a las dos escalas, pero hay pruebas de que puede ser posible, en algunos o en todos los casos, relacionar los dos ejes semánticos mediante una formulación adecuada y trabajar para continuar en este campo.

Los argumentos en favor de la escala continua son que algunas veces se justifica el tiempo suplementario de organización y de análisis debido a la fina discriminación que es tan posible como necesaria. Los argumentos en favor de una escala discreta son que no pueden lograrse resultados mejores, por razones que comprenden el hecho de que en la práctica, los no expertos no discriminan más que lo que permite la escala discreta.

Todos reconocen la necesidad de cierta forma de enclavamiento, y de que esto puede llevarse a cabo de diferentes formas. Algunos experimentadores sugieren que para las degradaciones que no se extienden a una amplia gama de valores, no se necesita ninguna imagen de referencia regular específica, porque la propia amplia gama normalizada de degradación hace que los observadores se orienten correctamente a sí mismo respecto a la escala. En cuanto a los experimentos que se refieren solamente a degradaciones muy pequeñas, debe proporcionarse una imagen de enclavamiento, pero no debe significarse como tal, debido a que esto haría demasiado artificial el medio ambiente de observación. El procedimiento de doble estímulo puede tener ventajas en las mediciones de degradaciones de pequeña magnitud tales como las que se producirán en sistemas futuros. Otros experimentadores sostienen que la utilización regular y significativa de una imagen de referencia (de alta calidad) ayuda a los observadores a orientarse a sí mismos respecto a la escala y que los resultados de los experimentos demuestran esto, particularmente cuando se trata de las degradaciones pequeñas. Otro método que se considera valioso es presentar dos imágenes de referencia (de alta y de baja calidad).

En cuanto al método de la UER, se encontró en un caso una nota para la referencia sólo de 4,63. En trabajos realizados por la Australian Broadcasting Commission [1981], hubo un caso en que la nota de la referencia sólo fue de 4,42. En evaluaciones de la SMPTE, utilizando NTSC y RGB como referencia, así como observadores procedentes de la industria de la radiodifusión, se advirtieron notas de la imagen de referencia de 4,7. Algunos autores creen que entre las razones de estos valores bajos puede figurar un control insuficiente del procedimiento o una calidad de referencia no coherente, en relación con el necesario enclavamiento directo. En tales casos, parece necesaria la corrección de la degradación residual, como en el caso de los procedimientos basados en un solo estímulo. Quizá conviniera realizar nuevos estudios relacionados con otros procedimientos.

En una evolución hacia la racionalización de los métodos, debe ser posible, en el próximo periodo de estudios del CCIR, obtener más elementos de los procedimientos juntos y confinar las alternativas sólo a determinadas partes. Es alentador observar que en una situación práctica reciente (el estudio de la relación entre la degradación y diferentes frecuencias de muestreo digital [CCIR, 1978-82e, f, g]) en la que se aplicaron cuidadosas reglas de procedimiento, virtualmente se lograron las mismas notas medias mediante pruebas totalmente independientes utilizando diferentes procedimientos (véanse los puntos 8.2 y 8.3 del apéndice a la Recomendación 500).

REFERENCIAS BIBLIOGRÁFICAS

ALLNATT, J. W. [1980] Subjective assessment method for television digital codecs, *Electron. Lett.*, **16**, 450-451.

ALLNATT, J. W. y CORBETT, J. M. [agosto de 1974] Comparisons of category scales employed for opinion rating. *Proc. IEE*, Vol. 121, **8**, 785-793.

AUSTRALIAN BROADCASTING COMMISSION [1981] Tests of subjective impairment due to random noise (se publicará ulteriormente).

CORBETT, J. M. [marzo de 1970] Effect of observer adaptation on the results of television quality-grading tests. *Proc. IEE*, Vol. 117, **3**, 512-514.

IBA [1981] Subjective assessment of television quality, experimental and development Report 114/81.

KRETZ, F. y SALLIO, P. [septiembre-octubre de 1981] Comparaison de plusieurs méthodes d'évaluation subjective de la qualité des images: rôle des images de référence, de l'ancrage et de l'echelle de notation. *Radiodif.-Télév.*, Vol. 4/5, **69**, 37-42.

MACDIARMID, I. F. y ALLNATT, J. W. [junio de 1978] Performance requirements for the transmission of the PAL coded signal. *Proc. IEE.*, Vol. 125, **6**, 571-580.

SALLIO, P. y KRETZ, F. [abril de 1982] A comparison of two methods for the subjective evaluation of television pictures. Representation of the results in common units. *EBU Rev. Tech.*, **192**, 59-69.

WHITE, T. A. y ALLNATT, J. W. [1980] Double-stimulus quality rating method for television digital codecs. *Electron. Lett.*, **16**, 714-716.

Documentos del CCIR

[1978-82]: **a.** 11/257 (Francia); **b.** 11/258 (Francia); **c.** 11/71 (Francia); **d.** 11/288 (Reino Unido); **e.** 11/285 (Reino Unido); **f.** 11/292 (Estados Unidos de América); **g.** 11/343 (Japón).

[1982-86]: **a.** 11/306 (Francia); **b.** 11/111 (Francia).

BIBLIOGRAFÍA

BENNETT, D. [octubre de 1981] SMPTE component-coded digital video picture quality assessments *SMPTEJ*, Vol. 90, **10**, 960-967.

BERNATH, K., KRETZ, F. y WOOD, D. [abril de 1981] The EBU method for organizing subjective tests of television picture quality, *EBU Rev. Tech.*, **186**, 66-75.

FISHER, R. A. y YATES, F. [1970] *Statistical Tables for Biological, Agricultural and Medical Research.* Oliver and Boyd, Edimburgo, Escocia, Reino Unido.

ISHIHARA, S. [1949] *Tests for Colour Blindness.* H. K. Lewis, London, Reino Unido.

KRETZ, F. [septiembre-octubre de 1981] Représentation unifiée des résultats d'essais subjectifs (correction pour dégradation résiduelle). *Radiodif.-Télév.*, Vol. 4/5, **69**, 43-44.

MICELI, S. y ORLANDO, A. [octubre de 1977] Sampling procedures and goodness of Fit. Proc. International Symposium on Measurement in Telecommunication (URSI), Lannion, Francia.

PROSSER, R. D., ALLNATT, J. W. y LEWIS, N. W. [marzo de 1964] Quality grading of impaired television pictures. *Proc. IEE*, Vol. 111, **3**, 491-502.

SALLIO, P. y KRETZ, F. [abril-mayo de 1978] Qualité subjective en télévision numérique. Première partie: méthodologie de son évaluation. *Radiodif.-Télév.*, Vol. 2/5, **52**, 13-19.

WHITE, T. A [1980] Transmission of alphanumerics by television: assessment of typescript by «experts». Proc. 9th International Symposium on Human Factors in Telecommunications, New Jersey, 27-34.

WHITE, T. A. [1981] Transmission of alphanumerics by television. *Displays*, **2**, 295-299.

WHITE, T. A. y REID, G. M. [agosto de 1981] Quality of PAL colour television pictures impaired by random noise: stability of subjective assessment, *Proc. IEE.*, Vol. 128, Parte F, **4**, 231-236.

Documentos del CCIR

[1974-78]: 11/360 (Francia).

[1978-82]: 11/17 (UER); 11/259 (Francia); 11/287 (Reino Unido); 11/309 (Italia); 11/312 (Francia); 11/313 (Francia); 11/331 (República Democrática Alemana); 11/357 (Italia).