Unión Internacional de Telecomunicaciones

UIT-T

SECTOR DE NORMALIZACIÓN
DE LAS TELECOMUNICACIONES
DE LA UIT

J.149

(03/2004)

SERIE J: REDES DE CABLE Y TRANSMISIÓN DE PROGRAMAS RADIOFÓNICOS Y TELEVISIVOS, Y DE OTRAS SEÑALES MULTIMEDIA

Mediciones de la calidad de servicio

Método para especificar la precisión y la calibración cruzada de métricas de la calidad de vídeo

Recomendación UIT-T J.149

UIT-T

Unión Internacional de Telecomunicaciones

Recomendación UIT-T J.149

Método para especificar la precisión y la calibración cruzada de métricas de la calidad de vídeo

Resumen

Las métricas de la calidad de vídeo se utilizan para proporcionar valores calculados que están estrechamente correlacionados con las evaluaciones subjetivas del observador. En esta Recomendación se describen métodos para ajustar las curvas de valores objetivos de VQM a los datos subjetivos con objeto de facilitar el cálculo de la precisión, se proporcionan un algoritmo para cuantificar la precisión de una determinada VQM, un método de cálculo simplificado del error cuadrático medio para cuantificar la precisión de una VQM cuando los datos subjetivos tienen aproximadamente la misma varianza en toda la escala VQM, y un método para representar gráficamente los errores de clasificación con objeto de determinar la frecuencia relativa de los casos de "vínculo falso", "diferenciación falsa", "clasificación falsa" y "decisión correcta" para una determinada VQM.

Orígenes

La Recomendación UIT-T J.149 fue aprobada el 15 de marzo de 2004 por la Comisión de Estudio 9 (2001-2004) del UIT-T por el procedimiento de la Recomendación UIT-T A.8.

PREFACIO

La Unión Internacional de Telecomunicaciones (UIT) es el organismo especializado de las Naciones Unidas en el campo de las telecomunicaciones y de las tecnologías de la información y la comunicación. El Sector de Normalización de las Telecomunicaciones de la UIT (UIT-T) es un órgano permanente de la UIT. Este órgano estudia los aspectos técnicos, de explotación y tarifarios y publica Recomendaciones sobre los mismos, con miras a la normalización de las telecomunicaciones en el plano mundial.

La Asamblea Mundial de Normalización de las Telecomunicaciones (AMNT), que se celebra cada cuatro años, establece los temas que han de estudiar las Comisiones de Estudio del UIT-T, que a su vez producen Recomendaciones sobre dichos temas.

La aprobación de Recomendaciones por los Miembros del UIT-T es el objeto del procedimiento establecido en la Resolución 1 de la AMNT.

En ciertos sectores de la tecnología de la información que corresponden a la esfera de competencia del UIT-T, se preparan las normas necesarias en colaboración con la ISO y la CEI.

NOTA

En esta Recomendación, la expresión "Administración" se utiliza para designar, en forma abreviada, tanto una administración de telecomunicaciones como una empresa de explotación reconocida de telecomunicaciones.

La observancia de esta Recomendación es voluntaria. Ahora bien, la Recomendación puede contener ciertas disposiciones obligatorias (para asegurar, por ejemplo, la aplicabilidad o la interoperabilidad), por lo que la observancia se consigue con el cumplimiento exacto y puntual de todas las disposiciones obligatorias. La obligatoriedad de un elemento preceptivo o requisito se expresa mediante las frases "tener que, haber de, hay que + infinitivo" o el verbo principal en tiempo futuro simple de mandato, en modo afirmativo o negativo. El hecho de que se utilice esta formulación no entraña que la observancia se imponga a ninguna de las partes.

PROPIEDAD INTELECTUAL

La UIT señala a la atención la posibilidad de que la utilización o aplicación de la presente Recomendación suponga el empleo de un derecho de propiedad intelectual reivindicado. La UIT no adopta ninguna posición en cuanto a la demostración, validez o aplicabilidad de los derechos de propiedad intelectual reivindicados, ya sea por los miembros de la UIT o por terceros ajenos al proceso de elaboración de Recomendaciones.

En la fecha de aprobación de la presente Recomendación, la UIT no ha recibido notificación de propiedad intelectual, protegida por patente, que puede ser necesaria para aplicar esta Recomendación. Sin embargo, debe señalarse a los usuarios que puede que esta información no se encuentre totalmente actualizada al respecto, por lo que se les insta encarecidamente a consultar la base de datos sobre patentes de la TSB en la dirección http://www.itu.int/ITU-T/ipr/.

ÍNDICE

Recomendación UIT-T J.149

Método para especificar la precisión y la calibración cruzada de métricas de la calidad de vídeo

1 Alcance

Las métricas de la calidad de vídeo se utilizan para proporcionar valores calculados que están estrechamente correlacionados con las evaluaciones subjetivas del observador. En esta Recomendación se describen:

a) métodos para ajustar las curvas de valores objetivos VQM a los datos subjetivos con objeto de facilitar el cálculo de la precisión y para producir una escala normalizada de valores objetivos que pueden utilizarse para la correlación cruzada entre diferentes VQM;

b) un algoritmo (basado en el análisis estadístico de los datos subjetivos) para cuantificar la precisión de una determinada VQM;

c) un método de cálculo simplificado del error cuadrático medio para cuantificar la precisión de una VQM cuando los datos subjetivos tienen aproximadamente la misma varianza en toda la escala VQM;

d) un método para la representación gráfica de errores de clasificación que sirve para determinar la frecuencia relativa de los casos de "vínculo falso", "diferenciación falsa", "clasificación falsa", y "decisión correcta" para una determinada VQM.

Los métodos especificados en esta Recomendación se basan en la evaluación objetiva y subjetiva de la componente vídeo tal como se define en la Rec. UIT-R BT.601, utilizando métodos como los que se describen en la Rec. UIT-R BT.500-11. La VQM utiliza un conjunto de datos que consiste en valores objetivos y notas medias subjetivas de diversas fuentes de vídeo en movimiento (SRC) procesadas por diferentes circuitos ficticios de referencia (HRC). Un ejemplo de un conjunto de datos de ese tipo figura en el Informe del material de referencia del UIT-T (véase el apéndice I).

Los métodos especificados en esta Recomendación se pueden aplicar directamente a estos conjuntos de datos. Para mediciones que no están incluidas en estos conjuntos de datos, los métodos especificados en esta Recomendación proporcionan una estimación razonable de la precisión y permiten una calibración cruzada para aplicaciones que pueden considerarse similares al conjunto de datos definidos, y que pertenecen al mismo campo.

Los métodos especificados en esta Recomendación pueden combinarse con otros cálculos estadísticos a fin de evaluar la utilidad de una VQM. En el apéndice I se explica la utilización de los métodos. Es indispensable disponer de un proceso de verificación completa realizado por laboratorios independientes competentes para que se pueda considerar la inclusión de una VQM en una parte normativa de una Recomendación del UIT-R.

NOTA – La estructura y contenido de la presente Recomendación se diseñaran para facilitar su utilización a quienes ya están familiarizados con el material inicial original; debido a ello, no se aplica el estilo tradicional de las recomendaciones del UIT-T.

2 Referencias informativas

– ANSI T1.801.01-1995*, *Digital Transport of Video Teleconferencing/Video Telephony Signals – Video Test Scenes for Subjective and Objective Performance Assessment*.

* Las normas T1 son mantenidas por ATIS desde noviembre de 2003.

– ANSI T1.801.02-1996, *Digital Transport of Video Teleconferencing/Video Telephony Signals – Performance Terms, Definitions and Examples*.

– ANSI T1.801.03-2003, *Digital Transport of One-Way Digital Signals – Parameters for Objective Performance Assessment*.

– IEEE Standard No. 205-2001, *Measurement of Luminance Signal Levels*.

– Material de referencia del UIT-T (2004), *Objective perceptual assessment of video quality: Full reference television* (www.itu.int/ITU-T/studygroups/com09/docs/tutorial_opavc.pdf).

– Recomendación UIT-R BT.500-11 (2002), *Metodología para la evaluación subjetiva de la calidad de las imágenes de televisión*.

– U.S. Standards Committee T1 Technical Report T1.TR.73-2001, *Video Normalization Methods Applicable to Objective Video Quality Metrics utilizing a Full Reference Technique*.

– U.S. Standards Committee T1 Technical Report T1.TR.74-2001, *Objective Video Quality Measurement Using Peak-Signal-to-Noise Ratio Full Reference Technique*.

– U.S. Standards Committee T1 Technical Report T1.TR.75-2001, *Objective Perceptual Video Quality Measurement Using a JND-Based Full Reference Technique*.

– U.S. Standards Committee T1 Technical Report T1.TR.77-2002, *Data and sample program code to be used with the method specified in T1.TR.72-2001 for the calculation of resolving power of the video quality metrics in T1.TR.74-2001 and T1.TR.75-2001*.

3 Abreviaturas

En esta Recomendación se utilizan las siguientes siglas:

FR-TV Televisión de referencia completa hace corresponder (*full reference television*)

HRC Circuito ficticio de referencia (*hypothetical reference circuit*)

RMSE Error cuadrátivo medio (*root mean squared error*)

SRC Fuente (*source*)

VQEG Grupo de expertos en calidad vídeo (*video quality experts group*)

VQM Métricos de la calidad vídeo (*video quality metrics*)

4 Precisión de una VQM

Para utilizar una métrica de la calidad de vídeo (VQM) objetiva, es indispensable saber si la diferencia de puntuaciones entre los dos vídeos procesados tiene significado desde el punto de vista estadístico. Por consiguiente, es necesario cuantificar la precisión (o la resolución) de la VQM. Un diagrama puede dar una idea de esta resolución: en abscisas se representa cada punto de la nota VQM de una determinada fuente de vídeo (SRC) y la distorsión (circuito ficticio de referencia, o HRC) y en ordenadas la nota subjetiva correspondiente a una determinada observación de SRC/HRC. Cada par SRC/HRC (correspondiente a una determinada nota VQM) contiene una distribución de las notas medias subjetivas S dadas por distintos observadores, que representa (aproximadamente) las probabilidades relativas de S para un determinado par SRC/HRC. La resolución de una VQM puede definirse como la diferencia entre dos valores VQM para los cuales las correspondientes distribuciones de notas subjetivas tienen medias que son diferentes entre sí desde el punto de vista estadístico (normalmente un nivel de significación de 0,95).

De acuerdo con esta descripción cualitativa, en esta cláusula se describen dos métricas de la resolución, siendo cada una de ellas útil en un contexto diferente. Las métricas se describen

en 4.3 y 4.4. Asimismo, en 4.5 se describe un método para evaluar las frecuencias de los diferentes tipos de error intrínsecos a la VQM. Como ejemplo de implementación de todos los métodos, en el apéndice II se proporciona el código fuente de un programa en MATLAB (The Mathworks, Inc., Natick, MA).

4.1 Nomenclatura y escalas combinadas

Supóngase que cada par de SRC/HRC en un conjunto de datos es una "situación", y sea N el número de situaciones en este conjunto de datos. La nota subjetiva correspondiente a la situación i y al observador l se indica como S_{il}, y la nota objetiva para la situación i se indica como O_i. La media de una variable, por ejemplo un observador, se indica mediante un punto en la posición de la variable. Por ejemplo, la nota media de opinión de una situación se indica como $S_{i\bullet}$. Se hará una evaluación de las estadísticas de la nota subjetiva correspondientes a cada par (i, j) de estas situaciones para determinar la significación de la diferencia de VQM y así conocer la resolución de esta diferencia de VQM en función del valor de VQM.

Antes de iniciar cualquier análisis estadístico, las notas medias de opinión objetivas originales $S_{i\bullet}$ se transforman linealmente al intervalo [0, 1], definido como la *escala común:* 0 indica que no hay distorsión y 1 representa la máxima distorsión. Si *mejor* representa el valor sin distorsión de una nota subjetiva original y *peor* representa la máxima distorsión en la escala sujetiva original, las notas $\hat{S}_{i\bullet}$ adaptadas a la escala vienen dadas por la expresión:

$$\hat{S}_{i\bullet} = \frac{S_{i\bullet} - mejor}{peor - mejor}$$

Seguidamente, las notas VQM se transforman a esta escala común como un subproducto del proceso de ajuste de las notas VQM a los datos subjetivos, que se tratará en la siguiente cláusula.

4.2 Ajuste de valores VQM a los datos subjetivos

Mediante el ajuste se eliminan las diferencias sistemáticas entre los datos VQM y los datos subjetivos (por ejemplo una variación cc) que no proporcionan información útil alguna sobre la calidad. Además, el ajuste de todos los VQM a una escala común proporciona un método para realizar la calibración cruzada de esas VQM.

El método más sencillo de ajuste de datos es la regresión y la correlación lineal. En el caso de las notas subjetivas de la calidad de vídeo este método quizá no sea el más adecuado. La experiencia con otros conjuntos de datos de la calidad de vídeo (véase el material de referencia del UIT-T) demuestran que el ajuste de la VQM a las notas subjetivas siempre es malo en los extremos de las gamas de valores. Este problema se puede mejorar mediante un algoritmo de ajuste que utiliza métodos no lineales aunque monótonos (que preserva el orden). Si se utiliza un modelo no lineal adecuado, los errores de ajuste de valores objetivos a subjetivos serán menores y prácticamente cero en el centro de la gama.

Es posible determinar los métodos no lineales para convertir efectivamente la escala VQM a la escala común [0, 1]. Además de mejorar el ajuste de los datos con una VQM, la curva de ajuste también ofrece una ventaja adicional con respecto al ajuste lineal correspondiente a la escala natural (es decir, la escala original de la VQM): la distribución de los errores por pasar de valores objetivos a subjetivos alrededor de la curva modelo ajustada es menos dependiente de las notas VQM. Evidentemente, es posible que la transformación no lineal no evite totalmente la influencia de las notas en los errores de pasar de valores objetivos a subjetivos. Para evitar totalmente esta influencia, la mejor solución habría sido registrar el error de la transformación de valores objetivos a subjetivos como una función del valor de VQM. Sin embargo, los conjuntos de datos son normalmente muy pequeños para clasificarlos por valor de VQM y que sigan siendo eficaces desde el punto de vista

estadístico. Por consiguiente, se calcula una especie de media a lo largo de la gama VQM como se indica en 4.3.

En la figura 1 se muestra el ajuste mejorado del modelo a los datos que se obtienen mediante la transformación de las notas objetivas utilizando una función de ajuste. Como puede observarse, además de mejorar el ajuste de los datos con VQM, la curva también ofrece otra ventaja con respecto al ajuste lineal propio de la escala natural: la distribución de los errores por pasar de valores del modelo a datos, sobre la curva de modelo ajustada, es menos dependiente de la nota VQM.

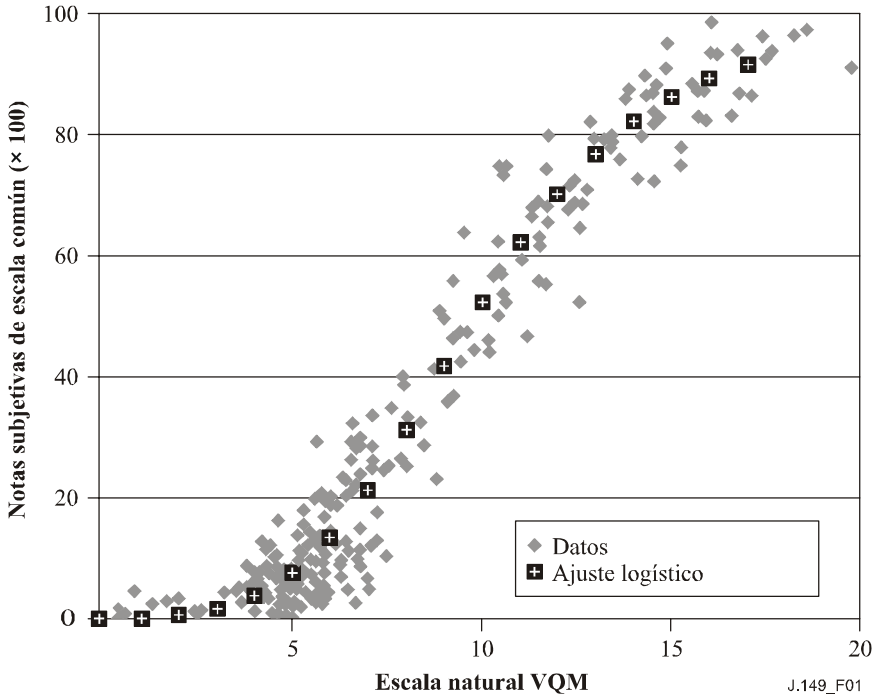

Figura 1 – Ajuste mejorado de datos a VQM mediante la correspondencia de VQM con una escala común

Sea O_i las notas objetivas originales (escala natural) y $\hat{O}_i$ las notas objetivas de la escala común. La relación entre ambas viene dada por la función de ajuste F (que depende de ciertos parámetros de ajuste). La función utilizada para ajustar los datos VQM objetivos (O_i) a los datos subjetivos adaptados a la escala ($\hat{S}_{i\bullet}$) debe tener los tres atributos siguientes:

a) un dominio específico de validez, que debería incluir la gama de datos VQM de todas las situaciones utilizadas para definir la métrica de la precisión;

b) una gama específica de validez, definida como la gama de notas en la escala común (una subgama de [0, 1]) con las que existe una relación de correspondencia de la función; y

c) ser monótona (la propiedad de ser siempre creciente o decreciente) a lo largo de todo el dominio de validez.

Evidentemente, la función de ajuste sería mucho más útil como herramienta de calibración cruzada si fuera monótona a lo largo de todo el dominio teórico de las notas VQM, si abarcara toda la escala común subjetiva de 0 a 1 y diera como resultado cero al aplicarla a la nota VQM que corresponde la secuencia de vídeo perfecta (sin degradaciones, y por tanto sin distorsión alguna). Sin embargo, este caso ideal no siempre puede lograrse para ciertas VQM y algunas familias de funciones utilizadas para realizar el ajuste.

Una de las posibles familias de funciones de ajuste es el conjunto de polinomios de orden M. Otra posibilidad es una función logística en la forma:

$$\hat{O}_i = a + b/\{1 + c(O_i + d)^e\}$$

siendo a, b, c, d, y e los parámetros de ajuste (véase el material de referencia del UIT-T). La tercera posibilidad es una función logística en la forma:

$$\hat{O}_i = a + (b-a)/\{1 + \exp[-c(O_i - d)]\}$$

siendo a, b, c, d parámetros de ajuste y $c > 0$[1]. Diremos para simplificar, que son las funciones lógicas Logística I y Logística II, respectivamente. El código MATLAB del apéndice II sólo contiene un ajuste por polinomios. En el apéndice III se describen posibles métodos de ajuste de datos utilizando funciones logísticas. La selección de la familia de funciones de ajuste (incluida la configuración a priori de algunos parámetros) depende de las notas asintóticas (peor y mejor) de la VQM en particular.

Del número de grados de libertad utilizado en el proceso de ajuste se indica mediante *D*. Por ejemplo, en un ajuste lineal $D = 2$, dado que se estima que hay dos parámetros de libertad en el procedimiento de ajuste. Se ha observado que la función de ajuste que transforma la VQM objetiva a la escala común facilita la comparación de dos VQM en la industria. El número de grados de libertad más adecuado 5, 4, 3 ó 2 depende del conjunto de datos en particular. Evítese sobreajustar los datos, porque habitualmente esto da lugar a resultados inestables y carentes de significado, y muchas veces impide la convergencia del algoritmo de ajuste.

Una vez transformada a la escala común, toda VQM puede calibrarse de manera cruzada con respecto a otra VQM a través de la escala común. La representación de la precisión de una VQM en la escala común facilita la comparación entre diversas VQM. Además, si la resolución en la escala común no varía mucho según la nota VQM en la cual se evalúa la resolución, la resolución puede mapearse en la escala natural utilizando la función logística inversa. Un ΔVQM de la escala común se traduce en una resolución ligada a la nota VQM en la escala natural. El cuadro o ecuación que proporciona esas resoluciones (una para cada nota VQM en la escala natural) tendrá un significado inmediato para los usuarios de la escala natural.

4.3 Métrica 1: Precisión de la VQM basada en la significación estadística

La nueva medida cuantitativa de la precisión VQM, denominada resolución, se define como el valor ΔVQM por encima del cual hay una diferencia estadística significativa entre las medias de las distribuciones de las notas subjetivas condicionales (normalmente a un nivel de significación de 0,95). Esta medida de la "barra de error" es necesaria para que los operadores de servicios vídeo puedan juzgar la significación de las fluctuaciones de VQM. Quizá no existan programas informáticos comerciales que calculen estadísticas de la resolución.

De todos los posibles métodos de evaluar la resolución de la VQM, se ha elegido la prueba-*t* de Student. Esta prueba se aplicó a la medición de todos los pares *i* y *j* de situaciones, para obtener el valor de ΔVQM (es decir la diferencia entre la nota VQM más grande y más pequeña de *i* y *j*) y la *significación* de la prueba-*t*. La significación es la probabilidad *p* de que, dados *i* y *j*, la nota VQM mayor corresponda a la situación que tiene la nota media subjetiva subyacente más alta. Así pues, *p* es la probabilidad de que la diferencia observada entre las medias de las muestras de las notas subjetivas correspondientes a *i* y *j* no correspondan a la media de una sola población, ni tampoco a medias de población que eran contrarias a las notas VQM conexas. Para obtener estos requisitos de

---

[1] Una versión modificada de esta función logística se utilizó en las cláusulas 6.2.3 y 6.2.4 del material de referencia del UIT-T. La modificación tiene en cuenta las diferencias en la varianza de las puntuaciones subjetivas.

ordenación, la prueba-*t* debe ser de una vía (*one-tailed*). En aras de la simplicidad, se realizó una prueba-*z* que corresponde aproximadamente a la prueba-*t*. Esta aproximación es bastante buena cuando el número de observadores es grande, lo cual era el caso para el conjunto de datos VQEG (material de referencia del UIT-T).

El análisis de la prueba de varianza (ANOVA) parece ser más adecuado que el método de la prueba-*t*. Sin embargo, aunque una sola aplicación de la ANOVA determina si existe una separación estadística entre el conjunto de categorías, se necesitan más comparaciones emparejadas para determinar las magnitudes y condiciones de diferencias significativas desde el punto de vista estadístico. Asimismo, en la ANOVA se sobreentiende que las varianzas de datos de la misma categoría son idénticas (lo cual puede no ser cierto). Por último, aunque existen muchos paquetes de programas que incluyen ANOVA, no siempre es fácil encontrar el paquete de programas informáticos adecuado (por ejemplo, no todas las rutinas ANOVA aceptan diferentes cantidades de datos en categorías distintas).

El algoritmo consta de los siguientes pasos:

Paso 1: Creación de una tabla de datos de entrada con N filas, donde cada fila representa una situación diferente (es decir, una fuente de vídeo y una distorsión diferente). Cada fila i consta de lo siguiente: el número de fuente, el número de distorsión, la nota VQM O_i, el número de respuestas N_i, la nota media subjetiva $S_{i\bullet}$, y la varianza de las muestras de las notas subjetivas V_i.

Paso 2: Transformación de las notas subjetivas $S_{i\bullet}$ a la escala común $\hat{S}_{i\bullet}$ como se describe en 4.1. También debe hacerse el cambio de escala de la varianza V_i de las notas subjetivas mediante la siguiente ecuación:

$$\hat{V}_i = \frac{V_i}{(peor - mejor)^2}$$

Obsérvese que la transformación de las notas subjetivas y de sus varianzas no es obligatoria, dado que no variará las estadísticas de *z* definidas a continuación, aunque sí podría variar el proceso de ajuste VQM. Seguidamente, se transforman las notas VQM O_i a la escala común utilizando la función de ajuste descrita en 4.2, y definida con mayor detalle en el apéndice III. El resultado del proceso de ajuste es un conjunto de notas VQM en la escala común $\hat{O}_i$. Se deben indicar los valores de los coeficientes utilizados en el ajuste y también el dominio VQM sobre el que se realizó el ajuste (dominio de validez).

Paso 3: Para cada par de situaciones distintas i y j ($i \neq j$), se utiliza una prueba-*z* de una sola vía para asignar la probabilidad de *significación* de la diferencia entre la VQM más grande y más pequeña ($\hat{O}_i$ y $\hat{O}_j$, respectivamente). La significación es la probabilidad que la nota VQM más grande corresponda a la situación que tiene la nota media subjetiva subyacente más alta. La nota *z* es:

$$z = \left(\hat{S}_{i\bullet} - \hat{S}_{j\bullet}\right) / \sqrt{\left(\hat{V}_i / N_i + \hat{V}_j / N_j\right)}$$

y la probabilidad de significación de la nota *z*, $p(z)$, es simplemente la función de distribución acumulativa de *z*:

$$p(z) = cdf(z) = (2\,\pi)^{-0,5} \int_{-\infty}^{z} exp(-z^2/2)\, dz$$

Paso 4: Creación de un diagrama de dispersión de $p(z)$ (ordenadas) con respecto a la nota ΔVQM (abscisas). Dadas N situaciones, se registra cada par (i, j) con $i > j$, se registra la diferencia de VQM $\hat{O}_i - \hat{O}_j$ en un vector de longitud $N(N-1)/2$ denominado **ΔVQM** (con índice k) y se registra la correspondiente nota *z* en un vector denominado $\boldsymbol{Z}$ de longitud $N(N-1)/2$ (con el mismo índice k). Conviene asegurarse de que ΔVQM(k) no será nunca negativa, lo cual puede garantizarse por

definición de la ordenación de los puntos de extremo *i* y *j*, que en otro caso es arbitraria. Para ello, si ΔVQM(*k*) es negativa, sustitúyase *Z*(*k*) por –*Z*(*k*) y ΔVQM(*k*) por –ΔVQM(*k*).

Paso 5: Considérense las 19 clasificaciones (indexadas mediante *m*) de ΔVQM que valen, cada una 1/10 de la gama total de ΔVQM. Las clasificaciones se solapan en un 50%. Se establece que ΔVQM_m es el punto central de cada clasificación, y p_m la media de *p*(*z*) para toda *z* en la clasificación *m*.

Paso 6: Se dibuja la curva que pasa a través de los puntos (ΔVQM_m, p_m), para producir un gráfico de *p* con respecto a ΔVQM. Obsérvese que *p* puede interpretarse como la probabilidad media de significación.

Paso 7: Se selecciona un umbral de probabilidad *p*, se dibuja una línea horizontal en el valor *p* de ordenadas y se determina el umbral ΔVQM en la intersección con la curva obtenida en el paso 6, definido como la precisión. Para una probabilidad media de significación de *p* o mayor, la ΔVQM debería sobrepasar este umbral. Los valores más comunes que se eligen para *p* son 0,68, 0,75, 0,90, y 0,95.

Una vez obtenido el valor de ΔVQM para un determinado *p*, puede utilizarse directamente en la escala común, lo cual sería adecuado para realizar la calibración cruzada descrita en la cláusula 5. Otra posibilidad, que se utiliza para otros propósitos, es reflejar inversamente este valor de ΔVQM en la escala natural para obtener la resolución en la escala natural R en función de la nota objetiva natural O:

$$R(O) = | F^{-1} [F(O) + \Delta VQM] - O|$$

siendo F la función de ajuste definida en 4.2. Para las funciones logísticas definidas en 4.2, la inversa de la logística I es:

$$F^{-1}(x) = [(1/c) (b/[x–a]) – 1]^{1/e} – d$$

y la inversa de la logística II es:

$$F^{-1}(x) = d – (1/c) \ln[(b–a)/(x–a) – 1]$$

Cuando |ΔVQM| << 1, R(O) viene dado aproximadamente por:

$$R(O) = |\Delta VQM / F'(O)|$$

siendo F'(O) la derivada de F con respecto a O. Esta aproximación suele ser suficiente para la mayoría de las aplicaciones.

NOTA – Para las funciones logísticas descritas en 4.2, la derivada de la logística I es:

$$F'(x) = –bce (x+d)^{e–1} /\{1 + c(x+d)^e\}^2$$

y la derivada de la función logística II es:

$$F'(x) = c (b–a) \exp[–c(x–d)]/\{1 + \exp[–c(x–d)]\}^2$$

4.4 Métrica 2: Cálculo del error cuadrático medio (RMSE) de VQM

Si los datos subjetivos tienen una varianza aproximadamente idéntica a lo largo de toda la escala VQM, es posible hacer una estimación común de la varianza o de la resolución. Como ejemplo, tomemos el error cuadrático medio (RMSE, *root-mean-squared error*). La hipótesis básica en el cálculo del RMSE de VQM es cuantificar el error cuadrático medio (MSE, *mean squared error*) entre los datos objetivos ajustados y sus correspondientes datos subjetivos. El RMSE de VQM entre los datos objetivos ajustados $\hat{O}_i$ y los datos subjetivos $\hat{S}_{i\bullet}$ adaptados a una escala se calcula mediante la expresión:

$$VQM\_RMSE = \sqrt{\frac{1}{N-D}\sum_{i=1}^{N}\left(\hat{O}_i - \hat{S}_{i\bullet}\right)^2}$$

siendo *N* el número total de situaciones (igual a *IJ*, donde *J* es el número de casos e *I* es el número de HCR), y *D* los grados de libertad utilizados en la curva de ajuste de los datos objetivos a los subjetivos realizados de acuerdo con 4.2. Es posible que no existan programas informáticos comerciales que implementen las estadísticas de error de clasificación.

4.5 Gráficos de clasificación

Una forma de evaluar la eficacia de una métrica de calidad de vídeo (VQM, *video quality metric*) es conocer los errores de clasificación. Hay un error de clasificación cuando los resultados de las pruebas subjetiva y la VQM son diferentes para un par de puntos de datos. En esta cláusula se discute el significado de los errores de clasificación, en cuanto a las representaciones gráficas de las notas z subjetivas con respecto a delta-VQM descritos en el texto principal. En la descripción que se hace a continuación se utiliza la escala común [0, 1] tanto para las notas subjetivas como para las objetivas. Según esta escala, el "0" indica que no hay distorsión y "1" representa la máxima distorsión.

Para una determinada prueba subjetiva es posible definir un umbral Δz, según el cual dos puntos de datos (A, B) son equivalentes desde el punto de vista estadístico o bien son diferentes[2]. Una vez realizado esto, los resultados de las pruebas subjetivas permiten clasificar cada par de puntos de datos (A, B) en una de las tres categorías siguientes:

	$\Delta z_{AB} < -\Delta z$	→ A es mejor que B	→ Bs
$-\Delta z \leq$	$\Delta z_{AB} \leq \Delta z$	→ A es igual a B	→ Es
$\Delta z <$	Δz_{AB}	→ A es peor que B	→ Ws

Las abreviaturas para estas tres categorías (Bs, Es, y Ws) indican subjetivamente mejor, subjetivamente equivalente y subjetivamente peor, respectivamente.

Considérese un umbral similar para los valores VQM, Δo:

	VQM(A) – VQM(B) < -Δo	→ A es mejor que B	→ Bo
–Δo ≤	VQM(A) – VQM(B) ≤ Δo	→ A es igual a B	→ Eo
Δo <	VQM(A) – VQM(B)	→ A es peor que B	→ Wo

Las abreviaturas para estas tres categorías (Bo, Eo, y Wo) indican objetivamente mejor, objetivamente equivalente y objetivamente peor, respectivamente.

Dado que de cada par de puntos de datos se somete a una triple clasificación mediante la prueba subjetiva y una triple clasificación independiente mediante la VQM, existen nueve resultados posibles. En la figura 2 se ilustran estos nueve espacios de resultados mediante las líneas de puntos en el espacio bidimensional de diferencias de notas subjetivas con respecto a diferencias VQM.

---

[2] Los puntos de datos A y B representan en realidad conjuntos de observaciones de dos pares SRC/HRC. Tal como se describió en el texto principal, la cantidad Δz_{AB} es la diferencia entre las medias de A y B $(\hat{S}_{A\bullet} - \hat{S}_{B\bullet})$, dividida por la desviación típica deducida $\sqrt{(\hat{V}_A / N_A + \hat{V}_B / N_B)}$, siendo $\hat{V}_A$ la varianza de las notas a partir de la situación A, y N_A el número de observaciones a partir de la situación A, etc.

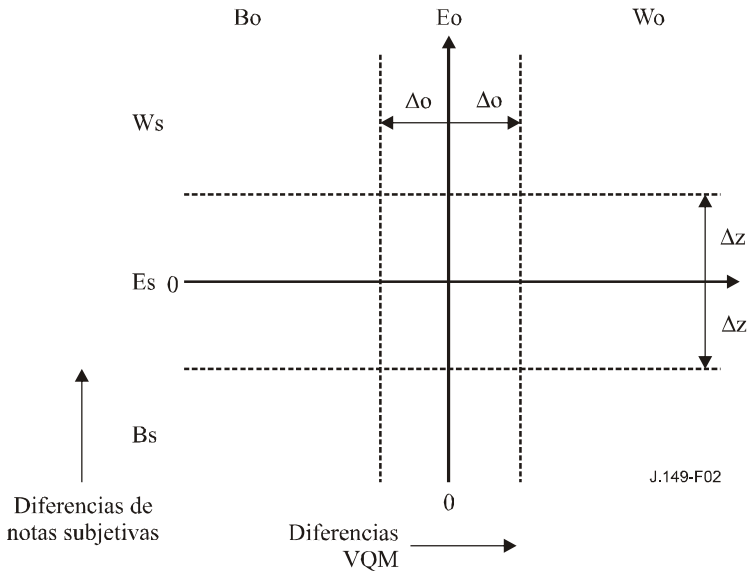

Figura 2 – Diagrama esquemático de clasificación

En el siguiente cuadro se da una descripción de cada uno de estos nueve resultados con miras a responder a la pregunta, ¿cómo se puede comparar la triple clasificación basada en VQM con la triple clasificación basada en las pruebas subjetivas?

	Bs	**Es**	**Ws**
Wo	Clasificación falsa	Diferenciación falsa	Decisión correcta
Eo	Vinculación falsa	Decisión correcta	Vinculación falsa
Bo	Decisión correcta	Diferenciación falsa	Clasificación falsa

Obsérvese que en tres de estos resultados la clasificación VQM es conforme con la clasificación de las pruebas subjetivas. Estos tres resultados se denominan "decisión correcta". Los seis resultados restantes corresponden a tres tipos de errores diferentes que pueden surgir cuando se utiliza una VQM. La vinculación falsa es probablemente el error menos grave: las pruebas subjetivas indican que dos puntos de datos son diferentes, pero la VQM indica que son iguales. La diferenciación falsa suele ser más grave: las pruebas subjetivas indican que dos puntos de datos son iguales, pero la VQM indica que son diferentes. Por lo general, la clasificación falsa sería el error más grave: según la prueba subjetiva A es mejor que B pero, según la VQM, B es mejor que A.

Dada una prueba subjetiva cualquiera y una VQM cualquiera, es posible construir todos los posibles pares distintos de puntos de datos y contar el número de pares que corresponden a cada una de las cuatro categorías de resultados distintas: decisión correcta, vinculación falsa, diferenciación falsa y clasificación falsa. Seguidamente se pueden normalizar por el número total de pares distintos y obtener las frecuencias relativas para estas cuatro categorías de resultados. En general, estos resultados serán funciones de Δz y de Δo. En el gráfico siguiente se muestran ejemplos de resultados de una VQM ficticia. Δz se seleccionó para tener una fiabilidad estimada del 95% en las clasificaciones subjetivas, y Δo es el parámetro libre en el eje x del gráfico.

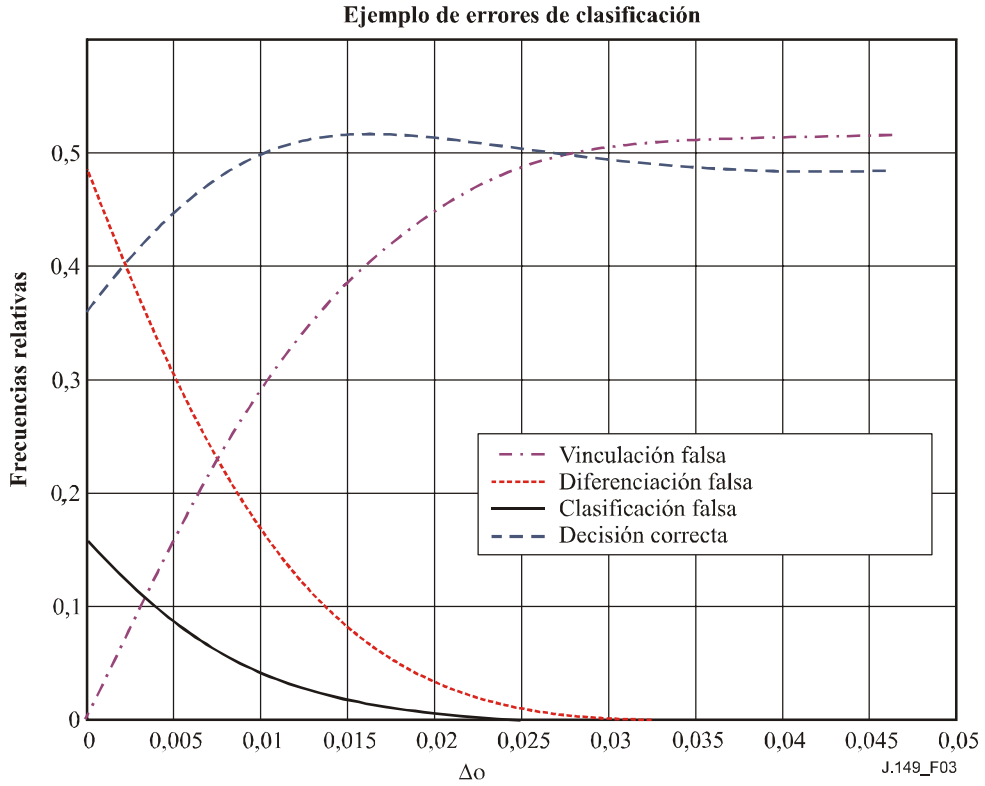

Figura 3 – Ejemplo de gráfico de frecuencias de errores de clasificación

Obsérvese que cuanto mayor es Δo, más pares de puntos de datos son equivalentes según la VQM. Es una forma de reducir el número de diferenciaciones falsas y clasificaciones falsas, pero aumenta la probabilidad de que se produzcan vinculaciones falsas. A medida que Δo se aproxima a 0,05, la tasa de vinculaciones falsas tiende a 0,52. En ese punto todos los pares son equivalentes según la VQM, y por esa razón la VQM es falsa el 52% de las veces y correcta el 48% de las veces. Esto es coherente con el hecho de que en esta prueba, el 48% de los pares de puntos de datos fueron declarados equivalentes mediante las pruebas subjetivas. Es posible utilizar un gráfico de este tipo para seleccionar el valor adecuado de Δo. Por ejemplo, es posible seleccionar Δo para maximizar la probabilidad de tomar decisiones correctas, o para minimizar la suma ponderada de las frecuencias relativas de error.

En el código mediante el cual se generó la figura anterior (parte del código MATLAB del apéndice II), el umbral utilizado para la prueba subjetiva es subj_th. El umbral utilizado para el ΔVQM, vqm_th, se dejó como un parámetro libre. El código representa gráficamente la frecuencia de los tres tipos de errores y los casos en que no hay errores, en función de vqm_th. Un posible valor óptimo de vqm_th es aquel que maximiza la frecuencia de casos en los que no hay error, o el que minimiza la suma de errores ponderada por gravedad. En general, es muy probable que las vinculaciones falsas sean el error menos grave, las definiciones falsas algo más graves y las clasificaciones falsas el peor tipo de error.

NOTA – Los nueve resultados y la cuadrícula de tres por tres en el espacio (ΔVQM, nota Z subjetiva) es la manera más natural de describir este análisis. Se supone que los valores de ΔVQM son bipolares, aunque el código ya haya tomado el valor absoluto de ΔVQM (y sustituido Z por −Z para todos los puntos cuyos valores de ΔVQM son negativos). Aunque esto no cambia el tratamiento matemático, la descripción más natural de la situación es, en este caso, seis resultados en una cuadrícula de 2 por 3. Dos resultados correctos (A mejor que B, y A peor que B) se han forzado uno sobre otro y hay todavía dos resultados de vinculación falsa, pero sólo uno de diferenciación falsa y uno de clasificación falsa.

5 Calibración cruzada de dos VQM

Para relacionar dos VQM se utiliza una transformación a la escala común descrita en las cláusulas 4.1-4.2[3]. Una vez se han transformado las dos VQM (por ejemplo VQM1 y VQM2) a la escala común (a través de un conjunto de datos subjetivos acordado) la transformación entre la métricas VQM1 y la VQM2 se realiza simplemente transformando primero la VQM a la escala común y después la transformación inversa de la escala común a la VQM2. Los modelos se comparan por referencia al mismo conjunto de datos. En caso de que no haya una correspondencia entre los dominios o gamas, la calibración cruzada debe declararse como indefinida. En esta Recomendación no se especifica ningún conjunto concreto de datos comunes.

---

[3] **CAVEAT. Las deducciones que se hagan de la calibración cruzada deben tomarse con precaución – por ejemplo, la calibración cruzada de dos VQM no significa que una de las VQM se pueda sustituir sin errores por la otra.** Una de las explicaciones de esta limitación es que el método de calibración cruzada presentado depende de los datos subjetivos concretos que definen la escala común. Podría argumentarse que no son necesarios datos subjetivos para la calibración cruzada y que se podrían conectar dos VQM directamente a través de sus resultados para un determinado conjunto de entradas (pares de vídeo de prueba y referencia). Sin embargo, independientemente del conjunto de entradas VQM seleccionadas para la calibración cruzada, las VQM pueden responder de manera diferente a algunos otros vídeos. Aún más importante, incluso con el mismo conjunto de entradas seleccionado, pueden darse cuatro entradas (1, 2, 3, 4) para las cuales las dos notas VQM varían en el mismo sentido pasando de 1 a 2, pero en sentidos opuestos pasando de 3 a 4. Este comportamiento es lo que hace que una VQM sea mejor que otra y no puede inferirse en ningún método de calibración cruzada.

Apéndice I

Aplicación de esta Recomendación para la evaluación y validación de posibles VQM

(Este apéndice no es parte integrante de esta Recomendación)

I.1 Elementos de una determinada VQM completa

Cada posible VQM debe ser validada de manera independiente y determinada completamente de modo que la pueda implementar directamente un profesional en la materia. La descripción de las nuevas VQM propuestas debe incluir tres conjuntos de datos distintos:

a) vectores de prueba para comprobar la implementación de la VQM, con las entradas vídeo y los resultados de la VQM;

b) datos de validación/precisión, que incluirán clasificaciones subjetivas y resultados del modelo (con una gama de calidad suficiente para que sea representativa de los vídeos transmitidos normalmente); y

c) datos relativos a otros métodos de evaluación, como por ejemplo el coeficiente de correlación lineal de Pearson entre las notas objetivas y subjetivas, la clasificación Spearman de la correlación de orden entre las notas objetivas y las subjetivas, y la relación Outlier. Por último, deben incluirse descripciones del alcance y las limitaciones, la precisión y el modelo de calibración cruzada como se describe en las siguientes cláusulas de esta Recomendación.

I.2 Alcance/limitaciones de una VQM

El alcance de una VQM puede incluir los siguientes elementos (se trata de una lista ilustrativa, que no pretende ser prescriptiva ni exhaustiva):

a) el tipo de contenido de la escena ("señal"), por ejemplo vídeo acelerado o en cámara lenta, en color o blanco y negro, entrelazado o progresivo;

b) el tipo y gravedad de las distorsiones ("ruido") causadas por las técnicas de codificación y las velocidades binarias (por ejemplo imagen borrosa, distorsión en los bloques);

c) las condiciones de visualización (incluidas la distancia de visualización, la iluminación del ambiente, los parámetros del monitor, por ejemplo el valor gamma, el brillo y los tipos de fósforo).

Cada VQM debe evaluarse cualitativamente en función del tipo de contenido de la escena, el tipo y gravedad de las distorsiones, y las condiciones de visualización en las que la VQM puede o no puede utilizarse eficazmente. Es importante enumerar todos los problemas conocidos y que no sean evidentes (por ejemplo las distorsiones de vídeo que incluyen cuadros perdidos) aunque se insiste en que el alcance/limitaciones descritas en esta cláusula no pretende ser exhaustivo.

En la descripción del alcance y limitaciones de la VQM se deben incluir cuatro cuadros. En los tres primeros se deben enumerar todas las distorsiones (circuitos ficticios de referencia o HRC) del conjunto de datos del Grupo de Expertos en calidad de vídeo (VQEG) y posiblemente otras, por ejemplo las siguientes:

a) un cuadro de los factores de prueba, las tecnologías de codificación y las aplicaciones para las que está comprobada la precisión de la VQM;

b) un cuadro de los factores de prueba, las tecnologías de codificación y las aplicaciones para las cuales se han hecho pruebas y *no* se ha obtenido la precisión de la VQM especificada en la cláusula 4; y

c) un cuadro de factores de prueba, tecnologías de codificación y aplicaciones conocidos para las cuales no se ha probado la VQM o para las que no se recomienda la utilización de esta VQM.

Además, debería incluirse:

d) un cuadro de las secuencias de prueba utilizadas para determinar los factores de prueba, las tecnologías de codificación y las aplicaciones para las cuales está demostrado que la VQM tiene precisión especificada en la cláusula 4.

Los siguientes cuadros corresponden a las pruebas de televisión de referencia completa de fase 1 VQEG (FR-TV). Dado que estos tres cuadros presentan de manera exhaustiva el conjunto de datos de fase 1 VQEG, el cuadro de muestras c no contiene ningún elemento. Los datos de fase 2 VQEG (material de referencia del UIT-T) no se han incluido porque los derechos de autor impiden poner el conjunto de datos a disposición general para la prueba de validación VQM.

Cuadro I.1 – Factores de prueba, tecnologías de codificación y aplicaciones para los cuales está demostrado que el método VQM tiene la precisión indicada

Velocidad binaria	Resolución	Método	Observaciones
2 Mbit/s	¾ resolución	mp@ml	Se reduce únicamente la resolución horizontal
2 Mbit/s	¾ resolución	sp@ml	
4,5 Mbit/s		mp@ml	
3 Mbit/s		mp@ml	
1,5 Mbit/s	CIF	H.263	
768 kbit/s	CIF	H.263	
4,5 Mbit/s		mp@ml	Señal compuesta NTSC y/o PAL
6 Mbit/s		mp@ml	
8 Mbit/s		mp@ml	Señal compuesta NTSC y/o PAL
8 & 4,5 Mbit/s		mp@ml	Dos códecs concatenados
19/PAL(NTSC)-19/PAL(NTSC)-12 Mbit/s		422p@ml	PAL o NTSC 3 generaciones
50-50-…-50 Mbit/s		422p@ml	7ª generación con desplazamiento/trama I
19-19-12 Mbit/s		422p@ml	3ª generación
ninguna		ninguno	Betacam de múltiple generación con abandono de trama (4 ó 5, compuesto/componente)

Cuadro I.2 – Factores de prueba, tecnologías de codificación y aplicaciones para los cuales no se ha obtenido la precisión indicada con el método VQM

Velocidad binaria	Resolución	Método	Observaciones
4,5 Mbit/s		mp@ml	con errores
3 Mbit/s		mp@ml	con errores

Cuadro I.3 – Secuencias de prueba utilizadas para determinar los factores de prueba, las tecnologías de codificación y las aplicaciones para los cuales está demostrado que la VQM tiene la precisión adecuada

Secuencia	Características
Balloon-pops	Película, color saturado, movimiento
NewYork 2	Efecto de enmascarado, movimiento
Mobile&Calendar	Disponible en ambos formatos, color, movimiento
Betes_pas_betes	Color, sintético, movimiento, corte de escena
Le_point	Color, transparencia, movimiento en todas las direcciones
Autumn_leaves	Color, paisaje, con zoom, movimiento de caída
Football	Color, movimiento
Sailboat	Casi detenida
Susie	Color de la piel
Tempete	Color, movimiento

Apéndice II

Código fuente MATLAB

(Este apéndice no es parte integrante de esta Recomendación)

A continuación se incluye la subrutina MATLAB denominada vqm_accuracy.m. Esta versión de la subrutina adapta los datos subjetivos a la escala [0, 1], aplica polinomios para ajustar los datos objetivos a los datos subjetivos a escala, calcula todas las métricas y representa gráficamente la frecuencia de casos "vinculación falsa", "diferenciación falsa", "clasificación falsa" y "decisión correcta" de la VQM. Es suficiente tener instalada la versión 5.3.1 de MATLAB (1999) y la librería de estadística y optimización que se vende por separado. Evidentemente, se puede elaborar un programa que no utilice ninguna de estas librerías. El código que se presenta a continuación sirve de ejemplo ilustrativo y no incluye todas las posibles opciones y funciones de ajuste.

Utilización: Para la VQM r0, escribir en la interfaz de usuario de matlab:

>load r0.dat

>vqm_accuracy(r0,-1,0,100,2)

Para la VQM r2, escribir:

>load r2.dat

>vqm_accuracy(r2,1,0,100,2)

r0.dat y r2.dat son ficheros de textos que contienen un subconjunto de los datos de línea 525 del VQEG. Cada línea en este fichero corresponde a una situación, y está formada por un número SRC, un número HRC, la nota VQM, el número de observaciones, la nota media subjetiva y la varianza de la nota subjetiva. Una vez cargados los dos ficheros r0 y r2.dat, se puede volver a ejecutar una de las dos subrutinas vqm_accuracy.

En el primer argumento de llamada de vqm_accuracy, r0 corresponde al modelo PSNR en TR A3, y r2 corresponde al modelo PQR en TR A4. El segundo argumento es 1 si la métrica objetiva indica menor calidad de la imagen cuando aumenta, y –1 en caso contrario. El tercer y cuarto argumentos son las puntuaciones nominales mejor y peor en la escala subjetiva natural. El último argumento es el orden del polinomio con el que se ajusta la VQM.

Código fuente:

```
function vqm_accuracy (data_in, vqm_sign, best, worst, order)
% MATLAB function vqm_accuracy (data_in, vqm_sign, best, worst, order)
%
% Each row of the input data matrix data_in must be organized as
% [src_id hrc_id vqm num_view mos variance], where
%
%  src_id is the scene number
%  hrc_id is the hypothetical reference circuit number
%  vqm is the video quality metric score for this src_id x hrc_id
%  num_view is the number of viewers that rated this src_id x hrc_id
%  mos is the mean opinion score of this src_id x hrc_id
%  variance is the variance of this src_id x hrc_id
%
%  The total number of src x hrc combinations is size(data_in,1).
%
% vqm_sign = 1 or -1 and gives the direction of vqm with respect to
%      the common subjective scale. For instance, since "0" is
%      no impairment and "1" is maximum impairment on the common
%      scale, vqm_sign would be -1 for PSNR since higher values
```

```
%      of PSNR imply better quality (i.e., this is opposite to
%      the common subjective scale).
%
% mos and variance will be linarly scaled such that
%  best is scaled to zero (i.e., the best subjective rating)
%  worst is scaled to one (i.e., the worst subjective rating)
%
% order is the order of the polynomial fit used to map the objective data
% to the scaled subjective data (e.g., order = 1 is a linear fit).
%

% Number of src x hrc combinations
num_comb = size(data_in,1);

% Pick off the vectors we will use from data_in
vqm = data_in(:,3);
num_view = data_in(:,4);
mos = data_in(:,5);
variance = data_in(:,6);

% Scale the subjective data for [0,1]
mos = (mos-best)./(worst-best);
variance = variance./((worst-best)^2);

% Use long format for more decimal places in printouts
format('long');

% Fit the objective data to the scaled subjective data.
% Following code implements monotonic polynomial fitting using optimization
% toolbox routine lsqlin.
%
% Create x and dx arrays. For the dx slope array (holds the derivatives of
% mos with respect to vqm), the vqm_sign specifies the direction of the slope
% that must not change over the vqm range.
x = ones(num_comb,1);
dx = zeros(num_comb,1);
for col = 1:order
  x = [x vqm.^col];
  dx = [dx col*vqm.^(col-1)];
end
% The lsqlin routine uses <= inequalities. Thus, if vqm_sign is -1 (negative
% slope), we are correct but if vqm_sign is +1 (positive slope), we must
% multiple each side by -1.
if (vqm_sign == 1)
  dx = -1*dx;
end
fit = lsqlin(x,mos,dx,zeros(num_comb,1));
fit = flipud(fit)' % organize this fit same as what is output by polyfit

% vqm fitted to mos
vqm_hat = polyval(fit,vqm);

% Perform the vqm RMSE calculation using vqm_hat.
vqm_rmse = (sum((vqm_hat-mos).^2)/(num_comb-(order+1)))^0.5

% Perform the vqm resolution measurement on both vqm and vqm_hat.
vqm_pairs = repmat(vqm,1,num_comb)-repmat(vqm',num_comb,1);
vqm_hat_pairs = repmat(vqm_hat,1,num_comb)-repmat(vqm_hat',num_comb,1);
mos_pairs = repmat(mos,1,num_comb)-repmat(mos',num_comb,1);
stand_err_diff = sqrt(repmat(variance./num_view,1,num_comb)+ ...
  repmat((variance./num_view)',num_comb,1));
z_pairs = mos_pairs./stand_err_diff;

% Include everything above the diagonal.
```

```
delta_vqm = [];
delta_vqm_hat = [];
z = [];
for col = 2:num_comb
  delta_vqm = [delta_vqm; vqm_pairs(1:col-1,col)];
  delta_vqm_hat = [delta_vqm_hat; vqm_hat_pairs(1:col-1,col)];
  z = [z; z_pairs(1:col-1,col)];
end

% Switch on z and delta_vqm for negative delta_vqm
z_vqm = z;
negs_vqm = find(delta_vqm < 0);
delta_vqm(negs_vqm) = -delta_vqm(negs_vqm);
z_vqm(negs_vqm) = -z_vqm(negs_vqm);

z_vqm_hat = z;
negs_vqm_hat = find(delta_vqm_hat <0);
delta_vqm_hat(negs_vqm_hat) = -delta_vqm_hat(negs_vqm_hat);
z_vqm_hat(negs_vqm_hat) = -z_vqm_hat(negs_vqm_hat);

% Plot scatter plot of z_vqm versus delta_vqm in figure 1.
% Plot scatter plot of z_vqm_hat versus delta_vqm_hat in figure 2.
figure(1)
plot(delta_vqm,z_vqm,'.','markersize',1)
set(gca,'LineWidth',1)
set(gca,'FontName','Ariel')
set(gca,'fontsize',12)
xlabel('Delta VQM')
ylabel('Subjective Z Score')
grid on
print -dpng figure1

figure(2)
plot(delta_vqm_hat,z_vqm_hat,'.','markersize',1)
set(gca,'LineWidth',1)
set(gca,'FontName','Ariel')
set(gca,'fontsize',12)
xlabel('Delta VQM Hat')
ylabel('Subjective Z Score')
grid on
print -dpng figure2

% Plot average confidence that vqm(2) is worse than vqm(1) in figure 3.
% Plot average confidence that vqm_hat(2) is worse than vqm_hat(1) in
% figure 4. These are the resolving power plots.
%
% One control parameter for delta_vqm resolution plot; number of vqm bins
% equally spaced from min(delta_vqm) to max(delta_vqm).
% Sliding neighbood filter with 50% overlap means that there will actually
% be vqm_bins*2-1 points on the delta_vqm resolution plot.
cdf_z_vqm = .5+erf(z_vqm/sqrt(2))/2;
cdf_z_vqm_hat = .5+erf(z_vqm_hat/sqrt(2))/2;

vqm_bins = 10; % How many bins to divide full vqm range for local averaging
vqm_low = min(delta_vqm); % lower limit on delta_vqm
vqm_high = max(delta_vqm); % upper limit on delta_vqm
vqm_step = (vqm_high-vqm_low)/vqm_bins; % size of delta_vqm bins

vqm_hat_low = min(delta_vqm_hat);
vqm_hat_high = max(delta_vqm_hat);
vqm_hat_step = (vqm_hat_high-vqm_hat_low)/vqm_bins;

% lower, upper, and center bin locations
low_limits = [vqm_low:vqm_step/2:vqm_high-vqm_step];
```

```
high_limits = [vqm_low+vqm_step:vqm_step/2:vqm_high];
centers = [vqm_low+vqm_step/2:vqm_step/2:vqm_high-vqm_step/2];

hat_low_limits = [vqm_hat_low:vqm_hat_step/2:vqm_hat_high-vqm_hat_step];
hat_high_limits = [vqm_hat_low+vqm_hat_step:vqm_hat_step/2:vqm_hat_high];
hat_centers = [vqm_hat_low+vqm_hat_step/2:vqm_hat_step/2: ...
    vqm_hat_high-vqm_hat_step/2];

mean_cdf_z_vqm = zeros(1,2*vqm_bins-1);
mean_cdf_z_vqm_hat = zeros(1,2*vqm_bins-1);
for i=1:2*vqm_bins-1
  in_bin = find(low_limits(i) <= delta_vqm & delta_vqm < high_limits(i));
  hat_in_bin = find(hat_low_limits(i) <= delta_vqm_hat & ...
    delta_vqm_hat < hat_high_limits(i));
  mean_cdf_z_vqm(i) = mean(cdf_z_vqm(in_bin));
  mean_cdf_z_vqm_hat(i) = mean(cdf_z_vqm_hat(hat_in_bin));
end

% The x-axis is vqm(2)-vqm(1). For figure 3 (the vqm plot), if vqm_sign is
% 1, then the Y-axis is the average confidence that vqm(2) is worse than
% vqm(1). On the other hand, if vqm_sign is -1, then the Y-axis is the
% average confidence that vqm(1) is worse than vqm(2). Figure 4 is the plot
% for vqm_hat, and since it always has the same sign as mos, the Y-axis is
% always the average confidence that vqm_hat(2) is worse than vqm_hat(1).
if (vqm_sign == 1)
  figure(3)
  % VQM resolving power
  plot(centers,mean_cdf_z_vqm)
  grid
  set(gca,'LineWidth',1)
  set(gca,'FontName','Ariel')
  set(gca,'fontsize',11)
  xlabel('VQM(2)-VQM(1)')
  ylabel('Average Confidence VQM(2) is worse than VQM(1)')
  print -dpng figure3
else
  figure(3)
  % VQM resolving power
  plot(centers,1-mean_cdf_z_vqm)
  grid
  set(gca,'LineWidth',1)
  set(gca,'FontName','Ariel')
  set(gca,'fontsize',11)
  xlabel('VQM(2)-VQM(1)')
  ylabel('Average Confidence VQM(1) is worse than VQM(2)')
  print -dpng figure3
end

figure(4)
% VQM Hat resolving power.
plot(hat_centers,mean_cdf_z_vqm_hat)
grid
set(gca,'LineWidth',1)
set(gca,'FontName','Ariel')
set(gca,'fontsize',11)
xlabel('VQM Hat(2) - VQM Hat(1)')
ylabel('Average Confidence VQM Hat(2) is worse than VQM Hat(1)')
print -dpng figure4

% This portion of the code calculates and plots the relative frequencies of
% three types of classification errors. A classification error is made when
% the subjective test and the VQM lead to different conclusions on a pair
% of data points.
%
```

```
% Background: For any subjective test, one must set a threshold that will
% determine when two results are statistically equivalent, and when they are
% statistically distinguishable. Then for each pair of data points (A,B),
% the subjective test can yield one of three possible outcomes: (1) A better
% than B, (2) A same as B, and (3) A worse than B.
%
% If we define a similar threshold for VQM values, we have the same
% situation. For each pair of data points, VQM can yield one of three
% possible outcomes: (1) A better than B, (2) A same as B, and (3) A worse
% than B. Since each pair of data points undergoes three-way classification
% by the subjective test and three-way classification by the VQM, there are
% nine possible outcomes. For three of these outcomes, the subjective test
% and the VQM agree. If we take the subjective test to be correct by
% definition, and the VQM to be under test, then we say that for these three
% outcomes, the VQM is correct. In two other cases the VQM has committed the
% "false-tie" error (subjective test says A better than B, or A worse than B,
% but VQM says A same as B). In two other cases the VQM has committed the
% "false differentiation" error (subjective test says A same as B, but VQM
% says A better than B, or A worse than B.) Finally, there are two cases
% where the VQM has performed a false ranking (subjective test says A better
% than B, or A worse than B, but VQM says the opposite.) Thus, all nine
% outcomes are accounted for. Note that a three by three grid in
% (delta_vqm, subjective Z score) space describing the above could be drawn.
%
% In the code below, the threshold used for the subjective test is subj_th.
% The threshold used for the delta VQM is vqm_th and this is left as a free
% parameter. The code plots the frequency of occurrence for the three
% different kinds of errors and for no error vs. vqm_th. An optimal value of
% vqm_th might be one that maximizes the frequency of occurrence of no error,
% or one that minimizes a cost-weighted sum of the errors. Note that in
% general, it is likely that false ties will be the least offensive error,
% false differentiations will be more offensive, and false rankings will be
% the worst sort of error.
%
% For more details, see S. Voran, "Techniques for Comparing Objective and
% Subjective Speech Quality Tests," Proceedings of the Speech Quality
% Assessment Workshop, Bochum, Germany, November 1994.
%
% Note: The nine outcomes and the three by three grid in (delta_vqm,
% subjective Z score) space is the most natural way to describe this
% analysis. This assumes bipolar values for delta_vqm. But the code has
% already taken the absolute value of delta_vqm (and replaced Z with -Z for
% all points with negative values of delta_vqm). This does not change the
% math, but the more natural description of the situation is now 6 outcomes
% and a 2 by 3 grid. Two correct outcomes (A better than B and A worse
% than B) have been folded on top of each other. There are still two false
% tie outcomes, but only one false differentiation outcome and one false
% ranking outcome.

% Figure 5 is the plot for vqm and figure 6 is the plot for vqm_hat.
subj_th = 1.6; % 95 percent confidence
num_th = 50; % number of delta_vqm thresholds to examine
vqm_th_list = [vqm_low:(vqm_high-vqm_low)/num_th:vqm_high];
vqm_hat_th_list = [vqm_hat_low:(vqm_hat_high-vqm_hat_low)/num_th: ...
    vqm_hat_high];
rel_freqs = zeros(vqm_bins+1,4);
rel_hat_freqs = zeros(vqm_bins+1,4);
for i = 1:num_th+1
  vqm_th = vqm_th_list(i);
  vqm_hat_th = vqm_hat_th_list(i);
  % Number of data points in the false tie region
  rel_freqs(i,1) = length(find((delta_vqm < vqm_th) & ...
    (subj_th <= abs(z_vqm))));
  rel_hat_freqs(i,1) = length(find((delta_vqm_hat < vqm_hat_th) & ...
```

```
    (subj_th <= abs(z_vqm_hat))));
  % Number of data points in the false differentiation region
  rel_freqs(i,2) = length(find((vqm_th <= delta_vqm) & ...
    (abs(z_vqm) < subj_th)));
  rel_hat_freqs(i,2) = length(find((vqm_hat_th <= delta_vqm_hat) & ...
    (abs(z_vqm_hat) < subj_th)));
  % Number of data points in the false ranking region
  if (vqm_sign == 1)
    rel_freqs(i,3) = length(find((vqm_th <= delta_vqm) & ...
      (z_vqm <= -subj_th)));
  else
    rel_freqs(i,3) = length(find((vqm_th <= delta_vqm) & ...
      (z_vqm >= subj_th)));
  end
  rel_hat_freqs(i,3) = length(find((vqm_hat_th <= delta_vqm_hat) & ...
    (z_vqm_hat <= -subj_th)));
end
% Normalize counts by total number of points to get relative frequencies
rel_freqs = rel_freqs/length(z_vqm);
rel_hat_freqs = rel_hat_freqs/length(z_vqm_hat);
% Calculate relative frequency of correctness
rel_freqs(:,4) = (1-sum(rel_freqs(:,1:3)'))';
rel_hat_freqs(:,4) = (1-sum(rel_hat_freqs(:,1:3)'))';

% Figure 5 is plot for vqm and figure 6 is plot for vqm_hat.
figure(5)
% VQM Subjective Classification Errors
plot(vqm_th_list,rel_freqs(:,1),'m-.', vqm_th_list,rel_freqs(:,2),'r:', ...
  vqm_th_list,rel_freqs(:,3),'k-',vqm_th_list,rel_freqs(:,4),'b--');
grid
set(gca,'LineWidth',1)
set(gca,'FontName','Ariel')
set(gca,'fontsize',12)
xlabel('Delta VQM Significance Threshold')
ylabel('Relative Frequencies')
legend('False Tie','False Differentiation','False Ranking','Correct Decision')
print -dpng figure5

figure(6)
% VQM Hat Subjective Classification Errors
plot(vqm_hat_th_list,rel_hat_freqs(:,1),'m-.', ...
  vqm_hat_th_list,rel_hat_freqs(:,2),'r:', ...
  vqm_hat_th_list,rel_hat_freqs(:,3),'k-', ...
  vqm_hat_th_list,rel_hat_freqs(:,4),'b--');
grid
set(gca,'LineWidth',1)
set(gca,'FontName','Ariel')
set(gca,'fontsize',12)
xlabel('Delta VQM Hat Significance Threshold')
ylabel('Relative Frequencies')
legend('False Tie','False Differentiation','False Ranking','Correct Decision')
print -dpng figure6
```

Apéndice III

Ajuste de datos a una escala común de VQM

(Este apéndice no es parte integrante de esta Recomendación)

Como se indicó en 4.2, los datos VQM objetivos (O_i) se hacen corresponder en un nuevo dominio $\hat{O}_i = \mathrm{F}(O_i)$. Este dominio se calcula realizando un ajuste de O_i a los datos subjetivos a escala ($\hat{S}_{i\bullet}$) mediante una familia de funciones F (con parámetros de ajuste) que son monótonas y tienen las propiedades de correspondencia de gama indicadas en 4.2. A continuación se indican tres opciones posibles para la familia de funciones F y la forma de ajustar los datos utilizando estas formas funcionales.

III.1 Polinomio de orden M

Un polinomio que se ajuste a un conjunto de puntos de datos no siempre es monótono. La librería de optimización MATLAB contiene una función lsqlin que garantiza la cualidad de monótona a lo largo de toda la extensión de los datos. Sin embargo, que sea monótona para el dominio de datos existente no implica que sea monótona para todo el dominio teórico (por ejemplo, de 0 a infinito).

III.2 Función logística I

El ajuste de los datos VQM objetivos (O_i) a los datos subjetivos a escala ($\hat{S}_{i\bullet}$) puede realizarse utilizando una función logística:

$$\hat{O}_i = \mathrm{F}(O_i) = a + b/\{1 + c(O_i + d)^e\}$$

siendo a, b, c, d, y e los parámetros de ajuste. La función de ajuste debe obtenerse mediante un ajuste no lineal por mínimos cuadrados[4]. La parte de la función que se utiliza es la parte monótona para $O > -d$ (de ahí el límite $d > -\min(O)$), y la forma de la curva s adecuada al ajuste de datos se garantiza mediante la restricción $e > 1$.

En ciertos casos, asintóticamente al menos, la puntuación perfecta en el modelo objetivo de escala natural se puede hacer corresponder a cero (la mejor nota en la escala subjetiva) y la peor nota objetiva posible en la escala natural se puede hacer corresponder a la peor nota subjetiva (la unidad, en la escala común). Por ejemplo, considérese el siguiente caso: *la mejor nota objetiva es cero, la peor nota objetiva es infinito*, en este ejemplo el cero se mapea en cero e infinito se mapea en 1, de modo que $a = 1$ y $b = -(1 + cd^e)$, por tanto:

$$\mathrm{F}(O_i) = 1 - (1 + cd^e)/\{1 + c(O_i + d)^e\}$$

El ajuste se realizará sobre c, d, e, siendo $d, e > 0$.

---

[4] El ajuste de la curva no lineal por mínimos cuadrados restringido puede realizarse utilizando la función MATLAB lsqcurvefit.

III.3 Función logística II

El ajuste de los datos VQM objetivos (O_i) a los datos subjetivos a escala ($\hat{S}_{i\bullet}$) también se puede realizar utilizando otra función logística:

$$\hat{O}_i = F(O_i) = a + (b-a)/\{1 + \exp[-c(O_i - d)]\}$$

siendo a, b, c, y d los parámetros de ajuste y c > 0 (lo que se garantiza mediante la definición c = |C| para C real). Análogamente a la función logística I, la función de ajuste se debe obtener mediante un ajuste no lineal por mínimos cuadrados[5].

Es posible utilizar esta optimización en el caso indicado en III.2: *mejor nota objetiva es cero, peor nota objetiva es infinito*. En este ejemplo cero se hace corresponder en cero, e infinito se hace corresponder en 1, de modo que: a = –exp[–cd] y b = –a exp[cd]. Por lo tanto:

$$F(O_i) = [1 - \exp(-cO_i)]/[1 + \exp\{c(d-O_i)\}]$$

La función logística II también es útil en el caso siguiente (que puede plantearse cuando O_i se expresa en coordinadas logarítmicas, por ejemplo decibelios): *mejor nota objetiva infinito, peor nota objetiva es menos infinito*. En este caso infinito se mapea en 0, y menos infinito se mapea en 1. Así pues b = 0, a = 1, y

$$F(O_i) = 1/[1 + \exp\{c(O_i - d)\}]$$

---

5 En la página 28 del Informe Final fase 1 VQEG (material de referencia del UIT-T) los valores iniciales para los parámetros eran a = nota subjetiva mínima, b = nota subjetiva máxima, c = 1, y d = nota media objetiva. También se utilizó una versión modificada de la logística II que tuviera en cuenta las diferencias en las varianzas de la puntuación subjetiva.

Bibliografía

[b-Voran] S. Voran, *Techniques for Comparing Objective and Subjective Speech Quality Tests*, Proceedings of the Speech Quality Assessment Workshop, Bochum, Germany, noviembre de 1994.

SERIES DE RECOMENDACIONES DEL UIT-T

Serie A	Organización del trabajo del UIT-T
Serie D	Principios generales de tarificación
Serie E	Explotación general de la red, servicio telefónico, explotación del servicio y factores humanos
Serie F	Servicios de telecomunicación no telefónicos
Serie G	Sistemas y medios de transmisión, sistemas y redes digitales
Serie H	Sistemas audiovisuales y multimedia
Serie I	Red digital de servicios integrados
Serie J	**Redes de cable y transmisión de programas radiofónicos y televisivos, y de otras señales multimedia**
Serie K	Protección contra las interferencias
Serie L	Construcción, instalación y protección de los cables y otros elementos de planta exterior
Serie M	Gestión de las telecomunicaciones, incluida la RGT y el mantenimiento de redes
Serie N	Mantenimiento: circuitos internacionales para transmisiones radiofónicas y de televisión
Serie O	Especificaciones de los aparatos de medida
Serie P	Terminales y métodos de evaluación subjetivos y objetivos
Serie Q	Conmutación y señalización
Serie R	Transmisión telegráfica
Serie S	Equipos terminales para servicios de telegrafía
Serie T	Terminales para servicios de telemática
Serie U	Conmutación telegráfica
Serie V	Comunicación de datos por la red telefónica
Serie X	Redes de datos, comunicaciones de sistemas abiertos y seguridad
Serie Y	Infraestructura mundial de la información, aspectos del protocolo Internet y Redes de la próxima generación
Serie Z	Lenguajes y aspectos generales de soporte lógico para sistemas de telecomunicación

Impreso en Suiza
Ginebra, 2009