RECOMENDACIÓN UIT-R BT.1676

Marco metodológico para la especificación de la precisión y la calibración mutua de las métricas de la calidad de la imagen

(Cuestión UIT-R 44/6)

(2004)

La Asamblea de Radiocomunicaciones de la UIT,

considerando

a) que la TV digital y la TVAD que utilizan tecnologías de reducción de la velocidad binaria, tales como la MPEG-2, la DV y otras han logrado una utilización generalizada;

b) que el Sector de Radiocomunicaciones se encarga de fijar las características generales de calidad de las cadenas de radiodifusión;

c) que las degradaciones de las imágenes de televisión pueden mostrar correlación con los aspectos medibles de las señales;

d) que la calidad general de la imagen está relacionada con la combinación de todas las degradaciones;

e) que en el caso de la TV digital, es necesario en particular evaluar la calidad de los métodos de reducción de la velocidad binaria, en términos de parámetros subjetivos y objetivos;

f) que en lo referente a los sistemas de televisión, se ha desarrollado una serie de parámetros objetivos de la calidad de la imagen, así como las correspondientes medidas de la calidad y métodos de supervisión para el entorno del estudio y en la radiodifusión;

g) que los métodos de referencia de medición objetiva de la calidad de la imagen son útiles al evaluar los sistemas de estudio y de radiodifusión;

h) que en las pruebas de validación de los métodos de medición objetiva de la calidad de la imagen se utilizan conjuntos de datos de materiales de prueba, notas subjetivas y valores objetivos;

j) que hay una serie de métricas de la calidad de la imagen (VQM) de referencia propuestas que pueden utilizarse para obtener evaluaciones objetivas de la calidad de la imagen;

k) que en la literatura se ha documentado una serie de métodos bien conocidos de evaluación estadística, los cuales pueden utilizarse para validar y comparar las VQM, basándose en los conjuntos de datos de materiales de prueba, las notas subjetivas y los valores objetivos;

l) que cuando se acepten una o más VQM como norma en las Recomendaciones de la UIT, seguirá existiendo la necesidad de estimar la precisión matemática (poder de resolución) de la VQM utilizada;

m) que la transcalibración de los métodos de referencia de medición objetiva de la calidad de la imagen basados en conjuntos de datos disponibles son importantes para el intercambio internacional de los resultados de las mediciones y la comprobación,

recomienda

1 que se utilicen los cálculos especificados en el Anexo 1 para estimar la precisión y la calibración mutua de las mediciones objetivas de la calidad de la imagen que utilizan el método de referencia;

2 que pueden utilizarse los cálculos especificados en el Anexo 1 como uno de los diversos métodos para determinar la precisión en la evaluación y validación de las diversas mediciones objetivas de la calidad de la imagen que utilizan el método de referencia.

Anexo 1

Método para especificar la precisión y la calibración mutua de las VQM

1 Alcance

El objetivo de las VQM es ofrecer valores calculados que guarden una estrecha correlación con las evaluaciones subjetivas del observador. Esta Recomendación incluye:

– Métodos de ajuste de curvas de valores objetivos de VQM a datos subjetivos, con el fin de facilitar el cálculo de la precisión y elaborar una escala de valores objetivos normalizados que pueda utilizarse en la correlación entre las distintas VQM.

– Un algoritmo (basado en los análisis estadísticos sobre los datos subjetivos) para cuantificar la precisión de una VQM determinada.

– Un cálculo simplificado del error medio cuadrático para cuantificar la precisión de una VQM cuando los datos subjetivos presentan una varianza prácticamente constante a lo largo de la escala VQM.

– Un método para representar los errores de clasificación, a fin de determinar las frecuencias relativas de los casos «vínculo falso», «diferenciación falsa», «clasificación falsa» y «decisión correcta» para una VQM determinada.

Los métodos especificados en esta Recomendación se basan en la evaluación objetiva y subjetiva de las componentes de vídeo, tal como se definen en la Recomendación UIT-R BT.601, utilizando métodos tales como los descritos en la Recomendación UIT-R BT.500 – Metodología para la evaluación subjetiva de la calidad de las imágenes de televisión. Un conjunto de datos para una VQM constará de valores objetivos y notas subjetivas medias correspondientes a diversas fuentes de imágenes en movimiento (SRC) procesadas por distintos circuitos ficticios de referencia (CFR). En el Informe Final de la Fase 1 del Grupo de Expertos en calidad de la imagen, Documento UIT-T COM 9-80-E – Final report from the video quality experts group on the validation of objective models of video quality assessment; figura un ejemplo de dicho conjunto de datos.

Los métodos especificados en esta Recomendación son aplicables directamente a un conjunto de datos definidos como el que se ha descrito. Para las mediciones que no forman parte específica del conjunto de datos, los métodos especificados en esta Recomendación ofrecen una estimación razonable de la precisión y la calibración mutua en las aplicaciones que pueden considerase similares a la del conjunto de datos definidos y dentro de su contexto.

Los métodos especificados en esta Recomendación pueden utilizarse en combinación con otros cálculos estadísticos al evaluar la utilidad de una VQM. El Apéndice 1 ofrece otras informaciones relativas a la utilización de los métodos. Para que se considere la inclusión de una VQM como parte normativa de una Recomendación del UIT-R se requiere un proceso de verificación completa realizado por laboratorios independientes adecuados.

2 Precisión de una VQM

Para utilizar una VQM objetiva, se debe saber si la diferencia de notas entre dos imágenes procesadas tiene significación estadística. Así pues, es preciso efectuar una cuantificación de la precisión (o potencia de resolución) de la VQM. Para visualizar esta potencia de resolución, conviene empezar con una representación de la dispersión en la que la abscisa de cada punto sea una nota de la VQM de una fuente de imagen particular (SRC) y la distorsión CFR, y la ordenada sea una nota subjetiva de una observación particular de la SRC/CFR. Cada combinación SRC/CFR (asociada a una nota de VQM particular) contiene una distribución de las notas medias subjetivas, *S*, basada en una serie de observadores, que representa (aproximadamente), las probabilidades relativas de *S* para la combinación SRC/CFR particular. La potencia de resolución de una VQM puede definirse como la diferencia entre dos valores VQM para los que la distribución correspondiente de notas subjetivas tenga valores medios que sean estadísticamente distintos entre sí (típicamente con un nivel de significación de 0,95%).

Dada esta imagen cualitativa, en este punto se describen dos métricas para la potencia de resolución, en la que cada una utiliza un contexto distinto. Las métricas se describen en los § 2.3 y 2.4. Además, en el § 2.5 se describe un método para evaluar las frecuencias de las distintas clases de errores cometidos por la VQM. Como ejemplo de implementación de todos los métodos, el Apéndice 2 ofrece un código fuente de computador en MATLAB (*The Mathworks, Inc., Natick, MA*).

2.1 Nomenclatura y escalas de coordinación

Llamando a cada combinación SRC/CFR en un conjunto de datos una «situación» y *N* el número de situaciones en este conjunto de datos, una nota subjetiva de la situación *i* y otorgada por el observador *l* se denominará S_{il}, y la nota objetiva de la situación *i* se denominará O_i. El promedio a lo largo de una variable tal como la del observador se indicará con un punto en la posición de dicha variable. Por ejemplo, la nota media de opinión de una situación se denominará $S_{i\bullet}$. Se han de evaluar las estadísticas de las notas subjetivas de cada par (*i, j*) de estas situaciones para entender el significado de la diferencia de las VQM y utilizarla a continuación para llegar a una potencia de resolución de la diferencia de las VQM, en función del valor de las VQM.

Antes de cualquier análisis estadístico se traspasan linealmente las notas medias de opinión subjetivas originales, $S_{i\bullet}$ al intervalo [0, 1], definido como escala común, representando el 0 la falta de degradación y el 1 la máxima degradación. Si *best* (*óptimo*) indica el valor sin degradación de la nota subjetiva original y *worst* (*peor*) representa la degradación máxima de la escala subjetiva original, las notas en escala, $\hat{S}_{i\bullet}$ vienen dadas por:

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

A continuación las notas de las VQM se transponen a su escala común como subproducto del proceso de ajuste de las notas de las VQM a los datos subjetivos, lo que se examinará en el punto siguiente.

2.2 Ajuste de valores de las VQM a datos subjetivos

El ajuste elimina las diferencias sistemáticas entre la VQM y los datos subjetivos (por ejemplo, la deriva continua) que no ofrece ninguna información útil de discriminación de calidad. Además, el ajuste de todas las VQM a una escala común dará un método para la calibración mutua de dichas VQM.

El método más sencillo de ajuste de datos es la correlación lineal y la regresión. Para las notas subjetivas de la calidad de la imagen, éste puede no ser el método óptimo. La experiencia con otros conjuntos de datos de calidad de vídeo indica sistemáticamente unos ajustes deficientes de la VQM a las notas subjetivas en los extremos de las gamas. Este problema puede mejorarse dejando que el algoritmo de ajuste utilice métodos no lineales si bien monótonos (que mantienen el orden). Si se utiliza un buen modelo no lineal, los errores de dato objetivo a nota subjetiva serán más pequeños y tendrán una tendencia central próxima a cero.

Los métodos no lineales pueden restringirse a la transformación efectiva de la escala de la VQM a la escala común [0, 1]. Además de mejorar el ajuste de los datos con una VQM, una curva de ajuste ofrece también una ventaja adicional respecto al ajuste de rectas que viene implícito en la escala nativa (es decir la escala original de la VQM): la distribución de los errores de dato objetivo a nota subjetiva alrededor de la curva del modelo de ajuste depende menos de las notas de las VQM. Evidentemente, la transformación no lineal puede no eliminar toda la dependencia respecto a las notas de los errores de dato objetivo a nota subjetiva. Para tener en cuenta la dependencia residual, habría sido útil en el caso ideal registrar el error entre dato objetivo y nota subjetiva en función del valor de la VQM. No obstante, los conjuntos de datos típicos son demasiado pequeños para dividirlos entre grupos de VQM en una forma estadísticamente potente. Por tanto, tal como se aclarará en el § 2.3, se calcula un tipo de promedio de la medición en la gama de la VQM.

La Fig. 1 muestra el ajuste mejorado del modelo a los datos obtenidos mediante la transformación de las notas objetivas utilizando una función de ajuste. Puede verse que, además de mejorar el ajuste de los datos con la VQM, la curva ofrece también una ventaja adicional respecto al ajuste de línea recta que implica la escala nativa: la distribución de los errores entre el modelo y los datos alrededor de la curva ajustada depende menos de las notas de la VQM.

A las notas objetivas originales (escala nativa) se le denominan O_i, y a las notas objetivas de la escala común $\hat{O}_i$. Una función de ajuste F (dependiente de algunos parámetros del ajuste) conecta las dos. La función utilizada para ajustar los datos objetivos de la VQM (O_i) a los datos subjetivos de la escala ($\hat{S}_{i\bullet}$) tiene que contar con los tres atributos siguientes:

– un dominio especificado de validez que debe incluir la gama de los datos VQM para todas las situaciones utilizadas en la definición de la métrica de precisión;

– una gama especificada de validez, definida como la gama de notas de escala común (una subgama de [0, 1]) con la que se corresponde la función; y

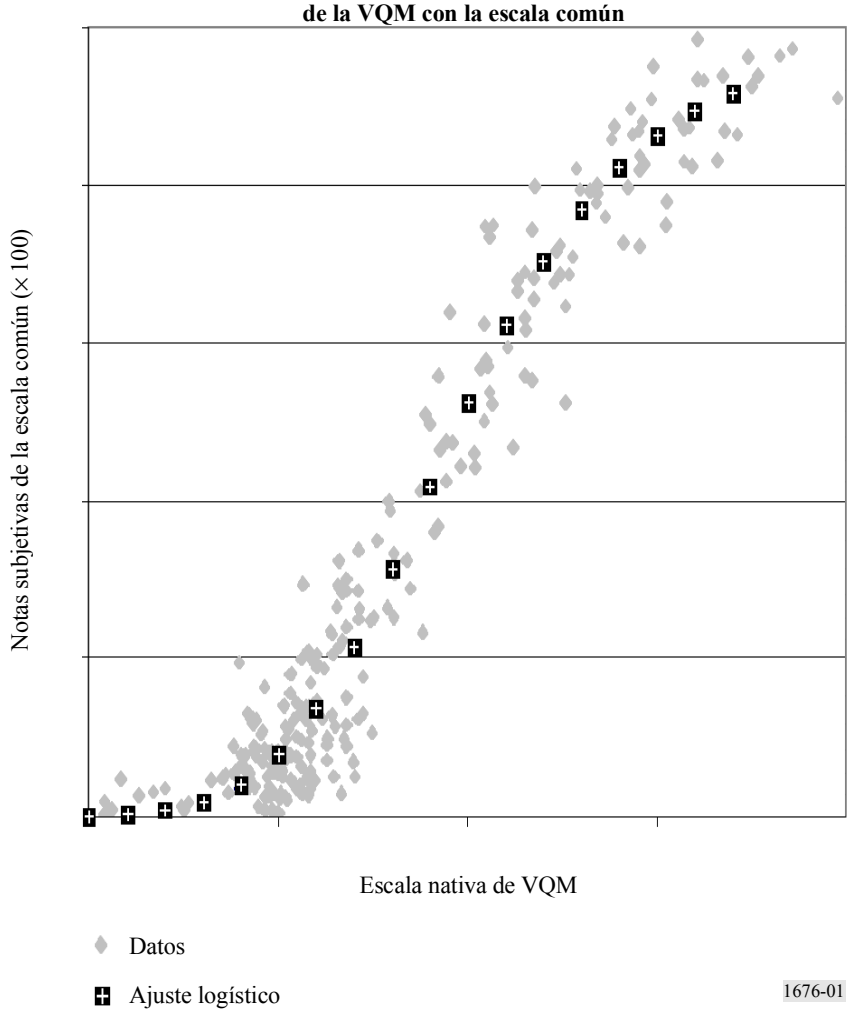
– monotonía (propiedad de ir estrictamente aumentando o estrictamente disminuyendo) a lo largo del dominio especificado de validez.

Evidentemente, la función de ajuste debe tener la máxima utilidad como instrumento de calibración mutua si es monótona a lo largo de todo el dominio teórico de las notas de la VQM, abarcase toda la escala común subjetiva de 0 hasta 1 e hiciese corresponder a cero la nota de la VQM que

corresponda a una secuencia de vídeo perfecta (sin degradaciones y por tanto con distorsión nula). No obstante, este caso ideal puede no ser alcanzable con ciertas VQM y familias de funciones utilizadas para el ajuste.

FIGURA 1

Ajuste mejorado de los datos a la VQM mediante la correspondencia de la VQM con la escala común



Una posible familia de funciones de ajuste es el conjunto de polinomios de orden M. Otra es una función logística que presenta la forma:

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

donde a, b, c, d y e son parámetros del ajuste. Una tercera posibilidad es una función logística que adopta la forma:

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

en la que a, b, c y d son parámetros del ajuste para $c > 0$. Por conveniencia, a estas formas logísticas se les denomina Logística I y Logística II, respectivamente. El código MATLAB del Apéndice 2 es únicamente un ejemplo de ajuste polinómico. El Apéndice 3 examina los posibles métodos de ajuste

de datos utilizando las funciones logísticas. La selección de una familia de funciones de ajuste (incluyendo una fijación a priori de algunos de los parámetros) depende de las notas asintóticas (la mejor y la peor) de la VQM particular.

El número de grados de libertad utilizados por el proceso de ajuste se denomina *D*. Por ejemplo, si se utiliza un ajuste lineal, $D = 2$ pues se estiman dos parámetros libres en el procedimiento de ajuste. La función de ajuste que transforma la VQM objetiva en la escala común se dice que facilita la comparación por la industria de dos VQM.

Una vez transformada a la escala común, toda VQM puede ser calibrada con otra VQM mediante la escala común. La representación de la precisión de una VQM en una escala común facilita las comparaciones entre las VQM. Además, suponiendo que la potencia de resolución en la escala común no varía mucho con la nota de la VQM a la que se evalúa la potencia de resolución, dicha potencia de resolución puede hacerse corresponder a través de la inversa de la opción logística con la escala nativa. En esta última, la ΔVQM de la escala común genera una potencia de resolución dependiente de la nota de la VQM. Un cuadro o una ecuación que dé dichas potencias de resolución (una para cada nota de VQM en escala nativa) tendrá un significado inmediato para los usuarios de la escala nativa.

2.3 MÉTRICA 1: Precisión de la VQM basada en el significado estadístico

Se define una nueva medida cuantitativa de la precisión de la VQM, denominada potencia de resolución, como el valor de la ΔVQM por encima de la cual las distribuciones condicionales de las notas subjetivas tienen significación estadísticamente distintas entre sí (típicamente con un nivel de significado de 0,95%). Dicha medida de «barra de error» es necesaria para que otros operadores del servicio de vídeo puedan juzgar el significado de las fluctuaciones de la VQM.

De los diversos posibles enfoques para evaluar la potencia de resolución de una VQM se eligió la prueba *t*-Student. Esta prueba se aplicó a la medición en todas las parejas de situaciones *i* y *j*. De la prueba resultan las ΔVQM (es decir, las diferencias entre las notas VQM mayor y menor de *i* y *j*) y la significación de la prueba *t*. Esta significación es la probabilidad *p* de que, dados *i* y *j*, la nota VQM mayor se asocie a la situación que tenga la mayor nota verdadera subjetiva media subyacente. Así pues, *p* es la probabilidad de que la diferencia observada en las medias muestreadas de las notas subjetivas de *i* y *j* no proceda de una única media de población, ni de medias de población que estén ordenadas de forma opuesta a las notas VQM asociadas. Para tener en cuenta este requisito de orden, la prueba *t* debe ser de una sola rama. Por simplicidad, la prueba *t* se aproximó mediante una prueba *z*. Esta aproximación es válida cuando el número de observadores es grande, como es el caso del conjunto de datos VQEG (Documento UIT-T COM 9-80-E).

Un análisis de la prueba de la varianza (ANOVA) puede parecer mejor que el método de la prueba *t*. No obstante, aunque una aplicación única de la prueba ANOVA determinará si existe una separación estadística entre un conjunto de categorías, se necesitan nuevas comparaciones por pares para determinar las magnitudes y condiciones de las diferencias estadísticamente significativas. Además, en el método ANOVA se supone la igualdad de las varianzas de los datos por categorías (lo que puede no ser cierto). Por último, aunque el método ANOVA está en muchos programas informáticos, la búsqueda del programa informático adecuado puede no ser fácil (por ejemplo, no todas las rutinas ANOVA aceptarán cantidades diferentes de datos en las distintas categorías).

Los algoritmos tienen los pasos siguientes:

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

Paso 2: Se traspasan las notas subjetivas $S_{i\bullet}$ a la escala común $\hat{S}_{i\bullet}$ como se describe en el § 2.1. La varianza V_i de las notas subjetivas debe también situarse en la escala correspondiente:

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

Véase que la transformación de las notas subjetivas y sus varianzas es facultativa. No cambiará la estadística z definida más adelante, pero puede cambiar el proceso de ajuste de las VQM. A continuación, se traspasan las notas VQM O_i a la escala común utilizando una función de ajuste, tal como se ha visto en el § 2.2 y se explica más detalladamente en el Apéndice 3. El resultado del proceso de ajuste es un conjunto de notas VQM de escala común $\hat{O}_i$. Se visualizan los valores de los coeficientes utilizados en el ajuste y también el dominio VQM en el que se efectúa 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 rama única a fin de asignar una probabilidad de significación a la diferencia entre las VQM mayor y menor ($\hat{O}_i$ y $\hat{O}_j$, respectivamente). El significado es la probabilidad de que la nota VQM mayor proceda de la situación con la nota subjetiva media verdadera subyacente mayor. La nota z es:

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

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

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

Paso 4: Se crea una representación de la dispersión de $p(z)$ (ordenada) en función de la nota ΔVQM (abscisa). 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 nota de z correspondiente en un vector denominado Z de longitud $N(N-1)/2$ (con el mismo índice k). Se desea asegurar que ΔVQM(k) nunca es negativa, lo que puede garantizarse mediante la definición del orden que, de no ser así sería arbitrario, de los puntos finales i y j. Para asegurar que así es, si ΔVQM(k) es negativa, se sustituye $Z(k)$ por $-Z(k)$ y ΔVQM(k) por $-$ΔVQM(k).

Paso 5: Se consideran 19 grupos (indexados por m) de ΔVQM, extendiéndose cada uno 1/10 de la gama total de ΔVQM. Los grupos se superponen el 50%. Se asocia $\Delta\mathrm{VQM}_m$ con el punto medio de cada grupo y se asocia p_m con la media de $p(z)$ para todas las z en un grupo m.

Paso 6: Se traza una curva por los puntos ($\Delta\mathrm{VQM}_m$, p_m), con la que se obtiene un gráfico de p en función de ΔVQM. Véase que p puede interpretarse como la probabilidad media de la significación.

Paso 7: Se selecciona una probabilidad umbral *p* y se traza una línea horizontal en el valor de la ordenada *p* hasta que intercepte la curva del Paso 6, determinando así el umbral ΔVQM que se define como la precisión. Para una probabilidad media de significación *p* o superior, la ΔVQM debe rebasar este umbral. Las opciones habituales de *p* son 0,68, 0,75, 0,90 y 0,95.

Tras hallar un valor de ΔVQM para una *p* determinada, se puede utilizar ésta directamente en una escala común – lo que sería adecuado para la calibración del Paso 6. Como alternativa, a otros efectos, hay la opción de la correspondencia inversa de este valor de ΔVQM a la escala nativa, para obtener una potencia de resolución en escala nativa, *R* en función de la nota objetiva nativa *O*:

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

donde *F* es la función de ajuste definida en el § 2.2. Para las funciones logísticas del § 2.2 la inversa de la Logística I es:

$$F^{-1}(x) = \left[(1/c)\,(b/[x-a]) - 1\right]^{1/e} - d$$

y la inversa de la Logística II es:

$$F^{-1}(x) = d - 1\,(1/c)\,\text{In}\left[(b-a)/(x-a) - 1\right]$$

Cuando $|\Delta\text{VQM}| << 1$, $R(O)$ puede aproximarse mediante:

$$RO = \left| \Delta\text{VQM} / F'(O) \right|$$

donde $F'(O)$ es la derivada de *F* respecto a *O*. Esta aproximación debe ser suficiente para la mayor parte de los casos.

NOTA 1 – Para las funciones logísticas del § 2.2 la derivada de la Logística I es:

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

y la derivada de la Logística II es:

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

2.4 MÉTRICA 2: Cálculo del error medio cuadrático (RMSE) de la VQM

Si los datos subjetivos tienen básicamente una varianza igual a lo largo de la escala VQM, puede ser adecuado efectuar una estimación agrupada de la varianza, o potencia de resolución. A modo de ejemplo, se elige el RMSE. La idea básica que encierra el cálculo del RMSE de la VQM es cuantificar el error de la media de los cuadrados (MSE) entre los datos objetivos ajustados y los

datos subjetivos correspondientes. El RMSE de la VQM entre los datos objetivos ajustados $\hat{O}_i$ y los datos subjetivos en escala $\hat{S}_{i\bullet}$ se calcula de la siguiente manera:

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

donde:

N: número total de situaciones (igual a IJ, donde J es el número de escenas e I es el número de CFR)

D: grados de libertad utilizados en el ajuste de las curvas objetiva a subjetiva efectuado en el § 2.2.

2.5 Representaciones de la clasificación

Los errores de clasificación son una forma de evaluar la eficacia de una VQM. Un error de clasificación se comete cuando la prueba subjetiva y la VQM dan lugar a conclusiones distintas en un par de puntos de datos. En este punto se examina el significado de los errores de clasificación, en términos de las representaciones de los datos subjetivos de z en función de la ΔVQM descrita en el texto principal. En la descripción que sigue, se utiliza la escala común [0, 1] para las notas subjetiva y objetiva. En este caso, «0» representa la ausencia de degradación y «1» representa la degradación máxima.

Para toda prueba subjetiva, se puede fijar un umbral Δz que define cuándo dos puntos de datos (A, B) son estadísticamente equivalentes y cuándo son estadísticas distinguibles[1]. Una vez efectuado esto, los resultados de la prueba subjetiva permiten situar 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*

$-\Delta z \le \Delta z_{AB} \le \Delta z$ → A es el mismo que B → *Es*

$\Delta z < \Delta z_{AB}$ → A es peor que B → *Ws*

Las abreviaturas para las tres categorías (*Bs*, *Es* y *Ws*) denotan subjetivamente mejor, subjetivamente equivalente y subjetivamente peor, respectivamente.

A continuación se considera un umbral similar para los valores VQM, Δo:

VQM (A) – VQM (B) $< -\Delta o$ → A es mejor que B → *Bo*

$-\Delta o \le$ VQM (A) – VQM (B) $\le \Delta o$ → A es el mismo que B → *Eo*

$\Delta o <$ VQM (A) – VQM (B) → A es peor que B → *Wo*

Las abreviaturas para las tres categorías (*Bo*, *Eo* y *Wo*) denotan objetivamente mejor, objetivamente equivalente y objetivamente peor, respectivamente.

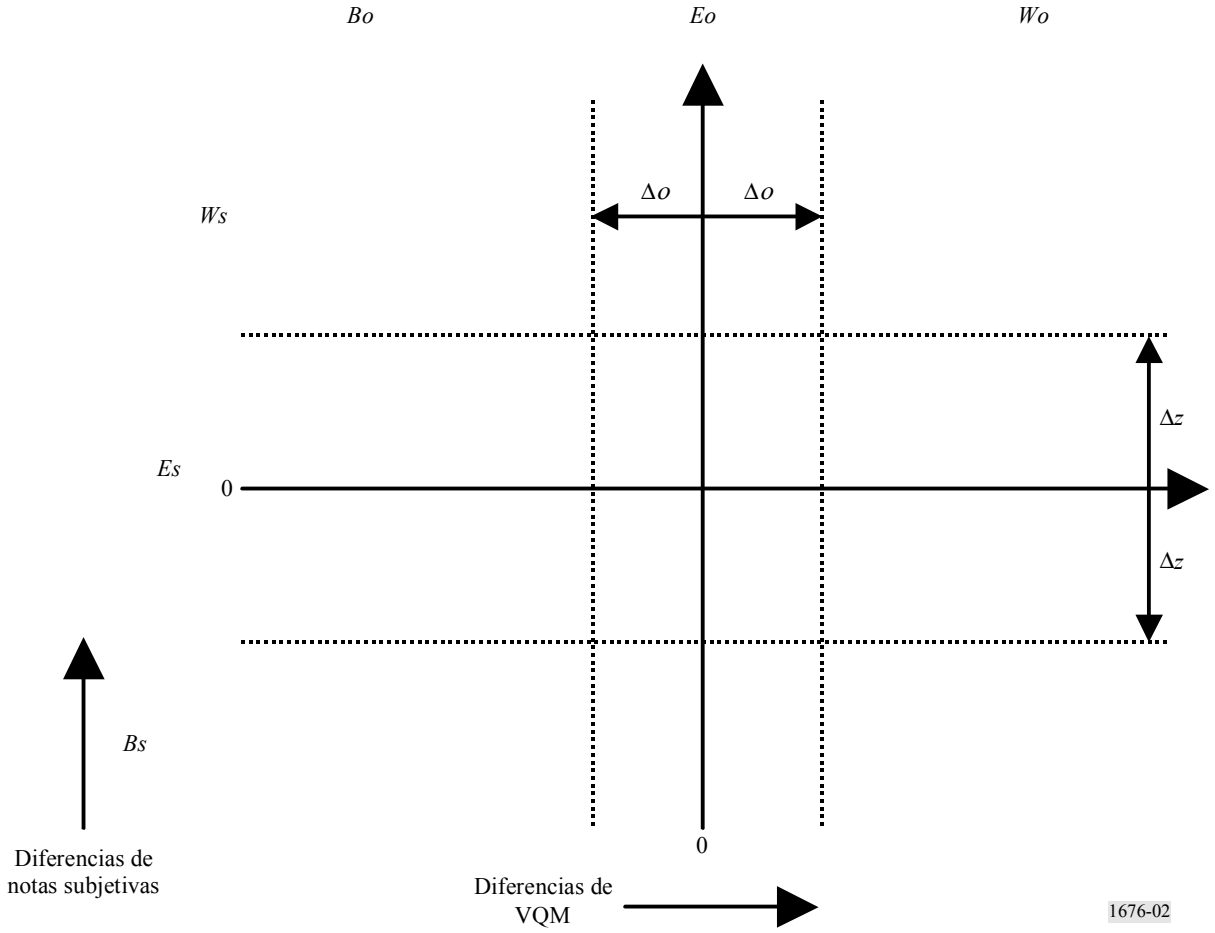
---

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

Como cada par de puntos de datos conlleva una clasificación triple por la prueba subjetiva y una clasificación triple separada por la VQM, hay nueve posibles resultados. Estos nueve espacios de resultado se ilustran en la Fig. 2 mediante las líneas de puntos del espacio bidimensional de la diferencia de notas subjetivas en función de la diferencia de VQM.

FIGURA 2

Diagrama de clasificación esquemática



En el Cuadro que sigue, se denomina a cada uno de estos resultados con una óptica encaminada a responder a la pregunta «¿Cómo se compara la clasificación triple basada en la VQM con la clasificación triple basada en la prueba subjetiva?».

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

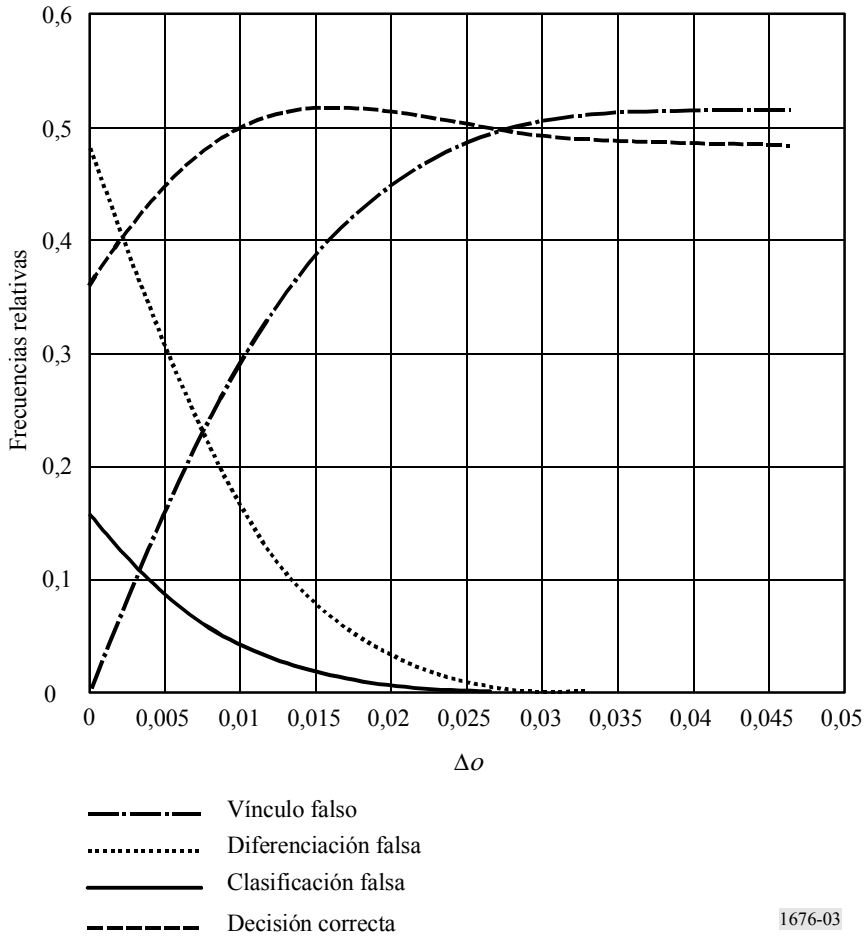
Véase que para tres de los resultados, la clasificación VQM concuerda con la clasificación de la prueba subjetiva. Estos tres resultados reciben el título de «decisión correcta». Los seis resultados restantes corresponden a tres tipos distintos de errores que pueden producirse cuando se utiliza

una VQM. El vínculo falso es probablemente el error menos grave. Se produce cuando la prueba subjetiva dice que dos puntos de datos son distintos, pero la VQM dice que son el mismo. Una diferenciación falsa suele ser más grave. Se produce cuando la prueba subjetiva dice que dos puntos de datos son el mismo, pero la VQM dice que son distintos. La clasificación falsa sería generalmente el error más grave. En la clasificación falsa, la prueba subjetiva dice que *A* es mejor que *B*, pero la VQM dice que *B* es mejor que *A*.

Para toda prueba subjetiva y toda VQM, podemos formar todos los posibles pares distintos de puntos de datos y contar el número de pares que pertenecen a cada una de las cuatro categorías distintas de resultados: decisión correcta, vínculo falso, diferenciación falsa y clasificación falsa. Se puede entonces normalizar respecto al número total de pares distintos e informar de frecuencias relativas para estas cuatro categorías de resultado. En general, estos resultados serán función de Δs y Δo. En la Fig. 3 se ofrecen ejemplos de resultados para una VQM ficticia. Se seleccionó Δz para obtener una confianza estimada del 95% en las clasificaciones subjetivas y Δo es el parámetro libre del eje *x* de la Figura.

FIGURA 3

Muestra de representación de las frecuencias del error de clasificación



Véase que a medida que Δo aumenta, la VQM indicará cada vez más pares de puntos de datos que son equivalentes. Ello reduce la aparición de diferenciaciones falsas y de clasificaciones falsas, pero aumenta la aparición de vínculos falsos. A medida que Δo se aproxima a 0,05, la tasa de vínculos falsos tiende hacia 0,52. En este punto, la VQM indica que todos los pares son equivalentes y al

hacerlo, la VQM se equivoca 52% de las veces y acierta el 48% de las veces. Ello es congruente con el hecho de que en esta prueba, se indicó que el 48% de los puntos de datos eran equivalentes según la prueba subjetiva. Se puede utilizar un gráfico como éste para seleccionar un valor adecuado de Δo. Por ejemplo, se puede seleccionar Δo para hacer máxima la probabilidad de efectuar decisiones correctas, o se puede seleccionar Δo para minimizar alguna suma ponderada de las frecuencias relativas de error.

En el código que generaba la Fig. 3 (parte del código MATLAB del Apéndice 2), el umbral utilizado para la prueba subjetiva es subj_th. El umbral utilizado para la ΔVQM, vqm_th, se deja como parámetro libre. El código representa la frecuencia de aparición de las tres clases distintas de errores y en caso de que no haya error, en función de vqm_th. Un valor óptimo de vqm_th sería el que hace máxima la frecuencia de aparición del caso sin errores o que minimice la suma ponderada de los errores. En general, es probable que los vínculos falsos sean el error menos grave, las diferenciaciones falsas sean el más grave y las clasificaciones falsas sean el tipo peor de error.

NOTA 1 – Los nueve resultados y la trama de tres por tres de espacio (ΔVQM, nota *Z* subjetiva) constituyen la forma más natural de describir este análisis. Se suponen en él valores bipolares de ΔVQM. Pero el código ya adoptado el valor absoluto de ΔVQM (y sustituido *Z* por –*Z* para todos los puntos con valores negativos de ΔVQM). Esto no cambia los aspectos matemáticos, pero la descripción más natural de la situación es ahora seis resultados y una trama de 2 por 3. Se han replegado dos resultados conexos (*A* mejor que *B* y *A* peor que *B*) en la parte superior. Sigue habiendo dos resultados de vínculo falso, pero sólo un resultado de diferenciación falsa y uno de clasificación falsa.

3 Calibración mutua de dos VQM

La necesidad de interrelacionar dos VQM se satisface mediante la transformación a una escala común descrita en los § 2.1-2.2[2]. Una vez transformadas dos VQM (por ejemplo, VQM1 y VQM2) a la escala común (por medio de un conjunto de datos subjetivos acordados), la transformación de VQM1 a VQM2 es simplemente la transformación directa de VQM1 a la escala común y a continuación la transformación inversa de la escala común a VQM2. Los modelos que se vayan a comparar tienen que hacer referencia a un conjunto de datos común. En los casos en que hay desadaptación de los dominios o las gamas de correspondencia, debe declararse que la calibración es indefinida. Esta Recomendación no especifica un conjunto de datos común particular.

---

[2] AVISO: Hay que actuar con cautela al efectuar inferencias de la calibración mutua – por ejemplo, la calibración de dos VQM no significa que una de ellas pueda sustituirse sin errores por la otra. Un motivo de esta limitación es que el actual método de calibración mutua depende de los datos subjetivos particulares que definen la escala común. Podría decirse que no se necesitan datos subjetivos para una calibración y que se pueden conectar dos VQM directamente mediante sus salidas, dado un conjunto particular de entradas (pares de vídeo de ensayo y referencia). No obstante, con independencia del conjunto de entradas VQM que se elijan para la calibración mutua, las VQM pueden responder distintamente a otras señales de vídeo. Lo que es más importante, incluso en el conjunto de entradas escogido, hay probablemente cuatro entradas (1, 2, 3, 4) tales que ambas notas VQM cambian en la misma dirección, pasando de 1 a 2, pero en sentidos opuestos pasando de 3 a 4. Dicho comportamiento es lo que hace que una VQM sea mejor que la otra y no se puede reflejar en ningún método de calibración mutua.

Apéndice 1
al Anexo 1

Aplicación de esta Recomendación a la evaluación y la validación de las VQM propuestas

(Informativo)

1 Elementos de la descripción completa de una VQM

Cada VQM posible debe validarse de forma independiente y describirse plenamente de forma que puedan aplicarla de inmediato los conocedores de la materia. La descripción de las VQM propuestas recientemente debe incluir tres conjuntos de datos distintos:

– vectores de prueba para verificar la implementación de la VQM, incluyendo entradas de vídeo y salidas resultantes de la VQM;

– datos de validación/precisión, incluyendo valoraciones subjetivas y resultados del modelo (que se extiendan en una gama de calidad suficiente de forma que sean representativos de las señales de vídeo transmitidas típicas);

– datos relativos a otros métodos de evaluación, tales como el coeficiente de correlación lineal de Pearson entre notas objetivas y subjetivas, la clasificación de Spearman de correlación de orden entre notas objetivas y subjetivas y la relación de Outlier.

Por último, debe haber descripciones del alcance y las limitaciones, la precisión y la calibración mutua del modelo, como las descritas en los puntos siguientes de esta Recomendación.

2 Alcance/limitaciones de una VQM

El alcance de una VQM puede incluir los elementos siguientes (lista ilustrativa que no pretende ser ni obligatoria ni total):

– el tipo de contenido de la escena (señal), por ejemplo movimiento alto/bajo, color respecto al blanco y negro, exploración entrelazada respecto a progresiva;

– el tipo y la intensidad de los efectos secundarios (ruido), inducidos por las técnicas de codificación y las velocidades binarias (por ejemplo, la borrosidad, la capacidad de bloqueo);

– las condiciones de observación (incluyendo la distancia de observación, la iluminación ambiente y los parámetros de la visualización, tales como la gama cromática, el brillo y los tipos de fósforos).

Cada VQM debe evaluarse cualitativamente en cuanto al tipo de contenido escénico, el tipo y la intensidad de los efectos secundarios y las condiciones de observación en las que puede o no puede funcionar de forma efectiva la VQM. Es importante enumerar las áreas problemáticas conocidas (tales como las de distorsiones de vídeo que incluyen pérdidas de cuadro) que de otra manera no serán obvias, si bien el apartado de alcance/limitaciones no pretende ser una lista completa.

Debe incluirse un conjunto de cuatro cuadros en la descripción del alcance y las limitaciones de la VQM. Los primeros tres cuadros deben enumerar todas las distorsiones (CFR) del conjunto de datos del grupo de expertos en calidad de vídeo (VQEG), y facultativamente otras, tales como las siguientes:

– un cuadro de factores de prueba, tecnologías de codificación y aplicaciones en las que la VQM han demostrado su precisión;

– un cuadro de factores de prueba, tecnologías de codificación y aplicaciones en las que se ha ensayado la VQM pero que *no* muestran la precisión especificada en el § 2;

– un cuadro de factores de pruebas conocidos, técnicas de codificación y aplicaciones (incluyendo todas las utilizadas por el VQEG) con las que no se ha ensayado la VQM o con las que ésta no se recomienda.

Además, debe haber un cuadro de secuencias de prueba utilizadas para determinar los factores de prueba, las tecnologías de codificación y las aplicaciones para las que la VQM ha mostrado la precisión especificada en el § 2.

A continuación se muestran ejemplos de los Cuadros.

CUADRO 1

a) Factores de prueba, tecnologías de codificación y aplicaciones con las que el método propuesto de la VQM ha mostrado la precisión especificada

Velocidad binaria	Resolución	Método	Comentarios
2 Mbit/s	3/4 resolución	mp@ml	Ésta es únicamente una reducción de la resolución horizontal
2 Mbit/s	3/4 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	NTSC compuesta y/o PAL
6 Mbit/s		mp@ml	
8 Mbit/s		mp@ml	NTSC compuesta 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 cuadro deriva/I
19-19-12 Mbit/s		422p@ml	3ª generación
No aplicable		No aplicable	Multigeneración Betacam con segregación (4 ó 5, compuesto/componente)

CUADRO 1 (*Fin*)

b) Factores de prueba, tecnologías de codificación y aplicaciones con las que el método propuesto de la VQM no ha mostrado la precisión especificada

Velocidad binaria	Resolución	Método	Comentarios
4,5 Mbit/s		mp@ml	Con errores
3 Mbit/s		mp@ml	Con errores

Estos Cuadros completan el conjunto de datos VQEG, de forma que el Cuadro 1c) no contendría ningún dato.

d) Secuencias de prueba utilizados para determinar factores de prueba, tecnologías de codificación y aplicaciones con las que la VQM ha mostrado la precisión especificada

Secuencia	Características
Balloon-pops (Globos que saltan)	Película, color saturado, movimiento
NewYork 2 (Nueva York 2)	Efecto de máscara, movimiento
Mobile & Calendar (Móvil y calendario)	Disponible en ambos formatos, color, movimiento
Betes_pas_betes	Color, sintético, movimiento, corte escénico
Le_point	Color, transparencia, movimiento en todas las direcciones
Autumn_leaves (Hojas de otoño)	Color, paisaje, aproximación, movimiento de caída de agua
Football (Fútbol)	Color, movimiento
Sailboat (Velero)	Casi estático
Susie	Color de piel
Tempête (Tempestad)	Color, movimiento

Apéndice 2
al Anexo 1

Código fuente MATLAB

(Informativo)

A continuación se indica una subrutina MATLAB denominada vqm_accuracy.m. Esta versión pone los datos subjetivos en escala [0, 1], aplica un ajuste polinómico de los datos objetivos a los datos subjetivos en escala, calcula todas las métricas y representa las frecuencias VQM de «vínculo falso», «diferenciación falsa», «clasificación falsa» y «decisión correcta». Es suficiente la versión 5.3.1 de MATLAB (1999) con los útiles de estadística y optimización de los que puede disponerse por separado. También puede desarrollarse un soporte lógico que no utilice ninguno de los útiles. Este código pretende mostrar un ejemplo ilustrativo y no incluye todas las opciones posibles y las funciones de ajuste.

Utilización: Al aparecer la indicación matlab, para el tipo r0 de VQM:

>load r0.dat

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

Para el tipo r2 de VQM:

>load r2.dat

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

En este caso, r0.dat y r2.dat son ficheros de texto que contienen un subconjunto de datos VQEG de 525 líneas. Cada línea de este fichero corresponde a una situación y comprende un número SRC, un número CFR, una nota VQM, un número de observaciones, la nota media subjetiva y la varianza de la nota subjetiva. Una vez cargados los ficheros r0 y r2.dat puede volverse a pasar cualquier patrón de vqm_accuracy.

En el primer argumento que aparece de vqm_accuracy, r0 corresponde al modelo PSNR de TR A3 y r2 corresponde al modelo PQR de TR A4. El segundo argumento es 1 si la métrica objetiva indica una peor calidad de la imagen cuando es mayor, y si no el argumento es –1. Los argumentos tercero y cuarto son las valoraciones nominales óptima y peor de la escala subjetiva nativa. El argumento final es el orden del polinomio al 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 3
al Anexo 1

Ajuste de datos a una escala común de VQM

(Informativo)

Tal como se ha examinado en el § 2.2 del texto principal, los datos VQM objetivos (O_i) se hacen corresponder con un nuevo dominio $\hat{O}_i = F(O_i)$. Este dominio se obtiene ajustando O_i a los datos subjetivos en escala ($\hat{S}_{i\bullet}$) mediante una familia de funciones F (con parámetros de ajuste) que tienen las propiedades de monotonía y correspondencia de gama señaladas en el § 5.2. A continuación se indican tres elecciones alternativas para la forma de F, junto con notas sobre el ajuste de datos que utilizan estas formas funcionales.

1 Polinomio de orden *M*

No está garantizado que un polinomio que se ajuste a un conjunto de puntos de datos no sea monótono. El útil de optimización MATLAB tiene una función, lsqlin que asegura la monotonía en toda la extensión de los datos. No obstante, la monotonía en el dominio de los datos existentes no asegura la monotonía en todo el dominio teórico (por ejemplo, de 0 a infinito).

2 Función Logística I

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

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

donde a, b, c, d y e son parámetros del ajuste. Las funciones del ajuste deben obtenerse a partir de mínimos cuadrados no lineales (véanse las notas MATLAB en el Informe Final VQEG[3]). La parte de la función que tiene que utilizarse es la parte monótona para $O > -d$ (y de ahí la restricción $d > -\text{mín}(O)$), y la forma de la curva s adecuada al ajuste de datos se asegura limitando $e > 1$.

En ciertos casos, al menos asintóticamente, puede hacerse que la nota perfecta del modelo de escala objetiva nativa se corresponda con cero (la nota óptima en la escala subjetiva) y que la nota peor objetiva de la escala nativa se corresponda con la nota peor subjetiva (la unidad en la escala común). Por ejemplo, considérese el caso siguiente: La nota objetiva óptima es cero, la nota objetiva peor es infinito. En este caso, cero se corresponde con cero e infinito se corresponde con uno, de forma que $a = 1$ y $b = -(1 + cd^e)$, y por tanto:

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

El ajuste tendrá lugar en c, d, e, a condición de que d, $e > 0$.

---

[3] El ajuste de curvas no lineal con mínimos cuadrados restringidos puede efectuarse mediante la función MATLAB lsqcurvefit.

3 Función Logística II

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

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

donde a, b, c y d, y $c > 0$ (se asegura definiendo $c = |\mathrm{C}|$ para C real). Como con la Logística I, la función de ajuste debe obtenerse a partir de mínimos cuadrados no lineales[4].

Se puede utilizar esta optimización en el caso señalado en el § 2: La nota objetiva óptima es cero, la nota objetiva peor es infinito. En este caso, cero se corresponde con cero e infinito se corresponde con 1, de forma que $a = -\exp[-cd]$ y $b = -a \exp[cd]$. Por tanto:

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

La función Logística II también es útil en el caso siguiente (que puede surgir cuando O_i se expresa en coordenadas logarítmicas, tales como decibelios): la nota objetiva óptima es infinito, la nota objetiva peor es infinito o negativo. En dicho caso, el infinito debe corresponderse con cero y el negativo infinito debe corresponderse con 1. Así pues, $b = 0$, $a = 1$, y:

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

---

[4] En la página 31 del Informe Final VQEG, los valores iniciales de los parámetros se eligen de forma que a = nota subjetiva máxima, b = nota subjetiva mínima, c = 1 y d = nota objetiva media.