RECOMMANDATION UIT-R BT.1676

Méthode de spécification de la précision des méthodes de mesure de la qualité vidéo et contre-étalonnage associé

(Question UIT-R 44/6)

(2004)

L'Assemblée des radiocommunications de l'UIT,

considérant

a) que les applications de télévision et de TVHD numériques utilisant des techniques de réduction du débit binaire telles que les normes MPEG-2 DV ou autres sont à présent très répandues;

b) que le Secteur des radiocommunications est chargé de définir les caractéristiques de qualité globale des chaînes de radiodiffusion;

c) qu'il est possible de montrer une corrélation entre les dégradations d'images de télévision et les caractéristiques mesurables des signaux;

d) que la qualité d'image globale est liée à la combinaison de toutes les dégradations;

e) que, dans le cas de la télévision numérique, il est notamment nécessaire d'évaluer la qualité des méthodes de réduction du débit binaire en ce qui concerne les paramètres tant objectifs que subjectifs;

f) que pour les systèmes de télévision, un certain nombre de paramètres de qualité d'image objectifs, ainsi que les méthodes de mesure et de contrôle associées, ont été développés pour les installations de studio et la radiodiffusion;

g) que les méthodes de mesure objective de la qualité d'image avec référence complète sont utiles pour évaluer les installations de studio et les systèmes de radiodiffusion;

h) que des jeux de données de test, des notes subjectives et des valeurs objectives sont utilisés pour les tests de validation de méthodes de mesure objective de la qualité d'image;

j) qu'un certain nombre de méthodes de mesure de la qualité vidéo (VQM, *video quality metrics*) avec référence complète ont été proposées, qui peuvent être utilisées pour fournir une évaluation objective de la qualité d'image;

k) qu'il existe dans la littérature technique un certain nombre de méthodes d'évaluation statistique bien connues qui peuvent être utilisées pour valider et comparer les méthodes VQM sur la base de jeux de données de test, de notes subjectives et de valeurs objectives;

l) que, après acceptation à titre normatif dans les Recommandations de l'UIT d'une ou plusieurs méthodes VQM, il sera toujours nécessaire d'évaluer la précision mathématique (le pouvoir de résolution) de la méthode VQM utilisée;

m) que le contre-étalonnage de méthodes de mesure objective de la qualité d'image avec référence intégrale sur la base des jeux de données disponibles est important pour l'échange international des résultats de mesure et de contrôle,

recommande

1 que les calculs spécifiés dans l'Annexe 1 soient utilisés pour évaluer la précision des méthodes de mesure objective de la qualité d'image, ainsi que leur contre-étalonnage, utilisant la méthode avec référence complète;

2 que les calculs spécifiés dans l'Annexe 1 puissent être utilisés comme l'une des nombreuses méthodes permettant de déterminer la précision de l'évaluation et de la validation de diverses méthodes de mesure objective de la qualité d'image utilisant la méthode avec référence complète.

Annexe 1

Méthode de spécification de la précision des méthodes VQM et contre-étalonnage associé

1 Domaine d'application

Les VQM sont censées fournir des valeurs calculées fortement corrélées aux évaluations subjectives des observateurs. La présente Recommandation contient:

– des méthodes d'ajustement des valeurs objectives VQM aux valeurs subjectives permettant de mieux déterminer la précision des calculs VQM et de générer une échelle de valeurs objectives normalisée qui peut être utilisée pour opérer une corrélation croisée entre différentes méthodes VQM;

– un algorithme (fondé sur une analyse statistique des données subjectives) permettant de quantifier la précision d'une méthode VQM;

– une méthode de calcul simplifiée de la valeur quadratique moyenne de l'erreur permettant de quantifier la précision d'une méthode VQM lorsque la variance des données subjectives est à peu près constante sur l'échelle VQM;

– une méthode de représentation graphique des erreurs de classification permettant de déterminer les fréquences relatives d'«équivalences erronées», de «différentiations erronées», de «classifications erronées» et de «décisions correctes» pour une méthode VQM donnée.

Les méthodes spécifiées dans la présente Recommandation sont fondées sur une évaluation objective et sur une évaluation subjective de la composante vidéo telle que celle-ci est définie dans la Recommandation UIT-R BT.601, en utilisant des méthodes telles que celles décrites dans la Recommandation UIT-R BT.500 – Méthodologie d'évaluation subjective de la qualité des images de télévision. Le jeu de données à considérer pour une méthode VQM comprendra les valeurs objectives et les notes subjectives moyennes correspondant à différentes sources d'images animées (SRC) associées à divers circuits fictifs de référence (HRC, *hypothetical reference circuits*). Un exemple de telles données figure dans le Document UIT-T COM 9-80-E – Final report from the video quality experts group on the validation of objective models of video quality assessment.

Les méthodes spécifiées dans la présente Recommandation sont directement applicables à un jeu de données bien défini. Pour les mesures qui n'appartiennent pas spécifiquement au jeu de données considéré, ces méthodes fournissent une estimation raisonnable de la précision et du contre-étalonnage pour les applications que l'on peut considérer comme similaire ou de même portée que celles associées au jeu de données précité.

Les méthodes spécifiées dans la présente Recommandation peuvent être associées à d'autres méthodes de calculs statistiques aux fins d'évaluation de l'utilité d'une méthode VQM. On trouvera dans l'Appendice 1 des informations relatives à l'utilisation de ces méthodes. Un processus complet de vérification par des laboratoires indépendants compétents est requis avant de pouvoir envisager l'inclusion à titre normatif d'une méthode VQM dans une Recommandation de l'UIT-R.

2 Précision d'une méthode VQM

Pour utiliser une méthode de VQM, on doit savoir si la différence de notation entre deux séquences vidéo traitées est importante d'un point de vue statistique. Il est donc nécessaire de quantifier la précision (ou le pouvoir de résolution) d'une méthode VQM. Pour visualiser ce pouvoir de résolution, il est utile dans un premier temps de tracer une courbe de dispersion dont l'abscisse est une note VQM relative à une source vidéo (SRC) particulière associée à une certaine distorsion HRC et l'ordonnée est une note subjective pour une observation particulière de cette paire SRC/HRC. A chaque paire SRC/HRC (à laquelle est associée une certaine note VQM) correspond une distribution de notes subjectives moyennes, *S*, attribuées par un certains nombre d'observateurs, distribution qui représente (approximativement) les probabilités relatives de *S* pour la paire SRC/HRC considérée. Le pouvoir de résolution d'une méthode VQM peut être défini comme la différence entre deux valeurs de mesures VQM pour laquelle les distributions correspondantes de notes subjectives présentent des moyennes statistiquement différentes (en général pour un intervalle de confiance à 95%).

Après cette description qualitative, nous présentons dans la suite du présent paragraphe deux méthodes de mesure du pouvoir de résolution dont chacune est adaptée à un contexte particulier. Ces méthodes sont décrites aux § 2.3 et 2.4. On décrit par ailleurs dans le § 2.5 une méthode permettant d'évaluer les fréquences d'occurrence de différents types d'erreurs engendrées par l'application d'une méthode VQM. Un code informatique source MATLAB (*The Mathworks, Inc., Natick, MA*) est donné dans l'Appendice 2 à titre d'exemple de mise en œuvre de toutes ces méthodes.

2.1 Nomenclature et échelles de coordonnées

Appelons «situation» chaque association source SRC/HRC d'un jeu de données et soit *N* le nombre de situations du jeu de données. On notera S_{il} la note subjective attribuée à une situation *i* par un observateur *l* et O_i la note objective attribuée à la situation *i*. La valeur moyenne calculée par rapport à une variable (l'observateur par exemple) sera signalée par un point placé à l'emplacement de cette variable. Ainsi, la note d'opinion moyenne attribuée à une situation sera notée $S_{i\bullet}$. Les valeurs statistiques de notes subjectives associées à chaque paire (*i*, *j*) d'une situation doivent être évaluées pour déterminer la signification de la différence entre mesures VQM puis utilisées pour parvenir à déterminer un pouvoir de résolution pour cette différence VQM, en fonction de la valeur VQM.

Avant toute analyse statistique, les notes d'opinion moyennes subjectives initiales $S_{i\bullet}$ sont «projetées» linéairement sur l'intervalle [0, 1], défini comme étant l'échelle commune de notation, 0 indiquant une absence de dégradation et 1 une dégradation maximale. Si *best* est la note subjective initiale en l'absence de dégradation et *worst* la note subjective initiale pour une dégradation maximale, les notes normalisées $\hat{S}_{i\bullet}$ sont données par:

$$\hat{S}_{i\bullet} = \frac{S_{i\bullet} - best}{word - best}$$

Les notes VQM sont ensuite «rapportées» à l'échelle commune de notation. Cette transformation est une conséquence du processus d'ajustement de ces notes aux données subjectives, ce qui sera examiné dans le prochain paragraphe.

2.2 Ajustement des valeurs VQM aux données subjectives

Le processus d'ajustement supprime les différences systématiques entre valeurs VQM et données subjectives (décalage de courant continu par exemple) qui n'apportent aucune information utile en termes de discrimination qualitative. De plus, l'application à toutes les méthodes VQM d'un processus d'ajustement pour que leurs valeurs appartiennent à une seule échelle commune de notation constituera une méthode de contre-étalonnage desdites méthodes.

La méthode d'ajustement de données la plus simple est celle de la corrélation linéaire et de la régression. Elle n'est peut-être pas la meilleure dans le cas de notes subjectives de la qualité vidéo. Pour d'autres jeux de données de qualité vidéo, l'expérience montre un ajustement de plus en plus mauvais entre les valeurs VQM et les notes subjectives aux extrémités d'intervalle. Ce problème peut être partiellement résolu en faisant intervenir dans l'algorithme d'ajustement des fonctions non linéaires mais toujours monotones (conservant l'ordre des séquences). Si un modèle non linéaire de bonne qualité est utilisé, les erreurs entre notes subjectives et notes objectives seront plus petites et auront tendance à converger vers zéro.

Un certain nombre de contraintes peuvent être appliquées aux méthodes non linéaires pour que celles-ci transforment effectivement l'échelle des valeurs VQM initiales en l'échelle commune de notation [0, 1]. Outre l'amélioration de l'ajustement des données aux valeurs VQM, l'utilisation d'une courbe d'ajustement offre un avantage supplémentaire par rapport à la mise en œuvre de l'ajustement linéaire induit par l'échelle de notation initiale (c'est-à-dire l'échelle de notation VQM initiale): la distribution des erreurs objectives/erreurs subjectives autour de la courbe de modélisation ajustée est moins dépendante des notes VQM. Bien sûr, cette transformation non linéaire peut ne pas supprimer toute dépendance des notes vis-à-vis des erreurs notes objectives/notes subjectives. Pour saisir cette dépendance résiduelle, il aurait été utile, idéalement, d'enregistrer les erreurs notes objectives/notes subjectives en fonction des valeurs VQM. Les jeux de données types sont toutefois trop petits pour être divisés en segments VQM d'une façon statistiquement robuste. C'est pourquoi, comme on le précisera au § 2.3, on calcule une sorte de mesure moyenne sur l'intervalle VQM.

La Fig. 1 montre l'amélioration de l'ajustement modèle/données induite par la transformation des notes objectives à l'aide de la fonction d'ajustement. On peut observer qu'outre l'amélioration de l'ajustement des données aux valeurs VQM, cette courbe offre un avantage supplémentaire par rapport à l'ajustement linéaire induit par l'échelle de notation initiale: la distribution des erreurs modèle/données subjectives autour de la courbe de modélisation ajustée est moins dépendante des notes VQM.

On désigne par O_i les notes objectives initiales (sur l'échelle de notation initiale) et par $\hat{O}_i$ les notes objectives sur l'échelle commune de notation. Une fonction d'ajustement F (qui dépend de certains paramètres d'ajustement) associe ces deux ensembles. La fonction utilisée pour ajuster les données VQM objectives (O_i) aux données subjectives normalisées ($\hat{S}_{i\bullet}$) doit satisfaire aux trois caractéristiques suivantes:

– un domaine de validité spécifié, qui devrait comprendre la plage des données VQM pour toutes les situations utilisées pour définir la mesure de la précision;

– un intervalle de validité spécifié, défini comme l'intervalle des notes sur l'échelle commune (il s'agit d'un sous-intervalle de [0, 1]) avec lequel la fonction met en correspondance le domaine de validité);

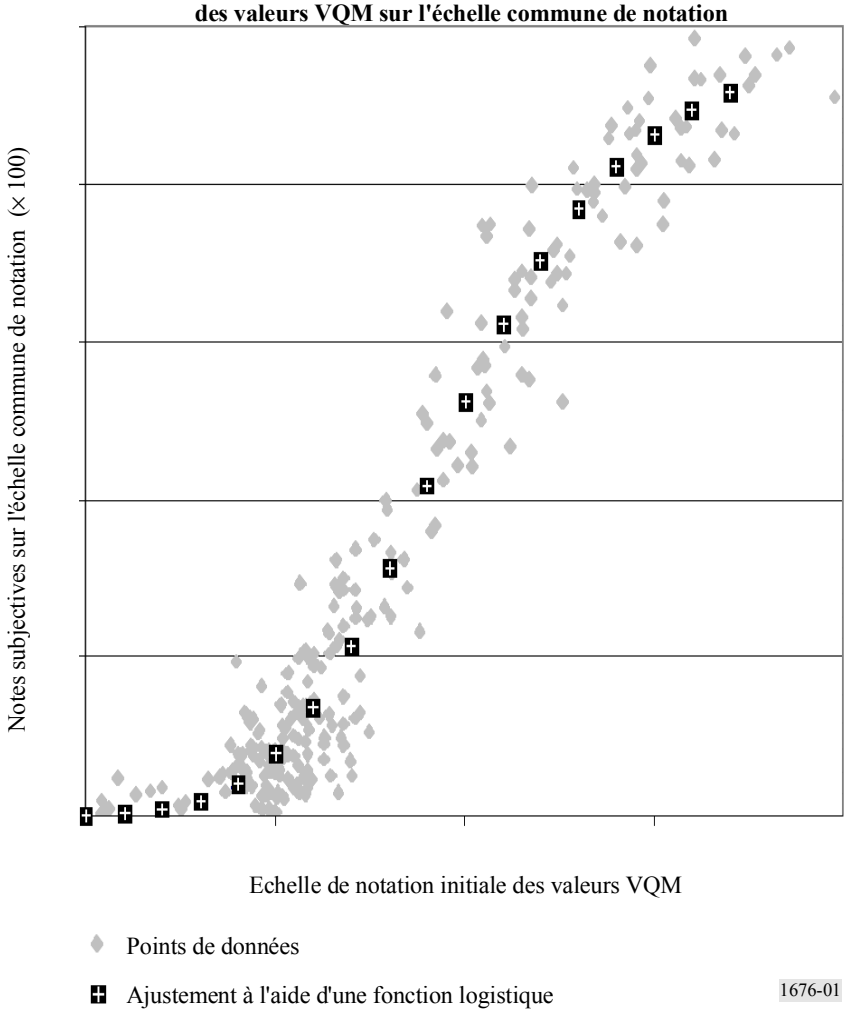
– un caractère monotone (fonction strictement croissante ou strictement décroissante) sur le domaine de validité spécifié.

Bien sûr, la fonction d'ajustement serait de la plus grande utilité en tant qu'instrument de contre-étalonnage si elle était monotone sur l'intégralité du domaine théorique des notes VQM, si elle couvrait l'intégralité de l'échelle commune subjective entre 0 et 1 et si elle faisait correspondre la

note VQM attribuée à une séquence vidéo parfaite (pas de dégradation, d'où une distorsion nulle) à la valeur zéro. Toutefois, cet idéal peut être hors de portée pour certaines méthodes VQM et familles de fonctions utilisées pour réaliser l'ajustement.

FIGURE 1

Amélioration de l'ajustement des données aux valeurs VQM grâce à la projection des valeurs VQM sur l'échelle commune de notation



Une famille de fonctions d'ajustement envisageable est l'ensemble des polynômes de degré *M*. Une autre famille est l'ensemble des fonctions logistiques de la forme:

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

où *a*, *b*, *c*, *d* et *e* sont les paramètres d'ajustement. Une troisième possibilité est offerte par la famille des fonctions logistiques de la forme:

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

où *a*, *b*, *c*, *d* sont les paramètres d'ajustement et $c > 0$. Pour plus de commodités, nous parlerons pour ces familles de fonctions logistiques respectivement des fonctions logistiques I et des fonctions logistiques II. Le code MATLAB de l'Appendice 2 n'utilise qu'une fonction d'ajustement polynomiale. On trouvera dans l'Appendice 3 l'examen de méthodes d'ajustement de données faisant

intervenir des fonctions logistiques. La sélection d'une famille de fonctions d'ajustement (y compris la détermination a priori de certains des paramètres) dépend des notes extrêmes (la meilleure et la plus mauvaise) pour la méthode VQM considérée.

D désigne le nombre de degrés de liberté utilisés par le processus d'ajustement. Pour un ajustement linéaire par exemple, $D = 2$ car deux paramètres indépendants sont évalués au cours de la procédure d'ajustement. La fonction d'ajustement qui projette les valeurs VQM objectives sur l'échelle commune de notation est censée faciliter les comparaisons entre deux méthodes VQM dans le monde industriel.

Une fois effectuée la projection vers l'échelle commune de notation, toute méthode VQM peut faire l'objet d'un contre-étalonnage avec n'importe quelle autre méthode VQM par l'intermédiaire de l'échelle commune de notation. Représenter la précision d'une méthode VQM sur une échelle commune facilite les comparaisons entre méthodes VQM. Par ailleurs, si l'on suppose qu'il ne varie pas beaucoup en fonction de la note VQM pour laquelle il est évalué, le pouvoir de résolution sur l'échelle commune de notation peut être projeté sur l'échelle de notation initiale par le biais de la fonction logistique inverse. A la variation ΔVQM sur l'échelle commune correspond sur l'échelle de notation initiale un pouvoir de résolution dépendant de la note VQM considérée. Un tableau ou une équation fournissant ces pouvoirs de résolution (un pour chaque note VQM sur l'échelle de notation initiale) aura une signification immédiate pour les utilisateurs de cette échelle.

2.3 MÉTHODE DE MESURE 1: précision de la méthode VQM fondée sur une signification statistique

Nous définissons une nouvelle mesure quantitative de la précision VQM, appelée pouvoir de résolution: il s'agit de la valeur ΔVQM au-dessus de laquelle les distributions conditionnelles de notes subjectives présentent des moyennes statistiquement différentes (généralement pour un intervalle de confiance à 95%). Une telle mesure de la «fourchette d'erreur» est nécessaire pour que les opérateurs de services vidéo puissent juger de la signification des fluctuations VQM.

Parmi plusieurs méthodes possibles d'évaluation du pouvoir de résolution d'une méthode VQM, le test *t* de Student a été choisi. Ce test a été appliqué pour les mesures de toutes les paires *i, j* de situations. Il permet d'obtenir la grandeur ΔVQM (c'est-à-dire la différence entre la meilleure note et la moins bonne note VQM pour les indices *i* et *j*) ainsi que la signification du test *t*. Le terme signification désigne la probabilité *p* que, pour *i* et *j* donnés, la note VQM la plus élevée soit associée à la situation pour laquelle la note subjective moyenne implicite réelle est la plus élevée. Ainsi, *p* est la probabilité que la différence observée des moyennes d'échantillons des notes subjectives pour les variables *i* et *j* ne soit due ni à une seule moyenne de population, ni à des moyennes de population classées dans un ordre opposé à celui des notes VQM associées. Pour saisir cette spécificité de classement, le test *t* doit être unidirectionnel. Pour plus de simplicité, il a été approximé par un test *z*. Il s'agit d'une bonne approximation lorsque le nombre d'observateurs est grand, ce qui était par exemple le cas pour le jeu de données du groupe VQEG (Document UIT-T COM9-80-E).

L'utilisation d'un test d'analyse de variance (ANOVA, *analysis of variance*) semble être plus appropriée que l'application de la méthode du test *t*. Toutefois, bien qu'une seule application de la méthode ANOVA permette de déterminer si une séparation statistique existe entre plusieurs catégories, d'autres comparaisons par paires sont nécessaires pour déterminer l'amplitude et les conditions d'apparition de différences statistiquement significatives. Par ailleurs, l'analyse ANOVA suppose que les variances associées à différentes catégories de données sont égales (ce qui peut ne pas être vrai). Enfin, bien que l'application ANOVA existe dans de nombreux prologiciels, choisir celui qui convient peut ne pas être facile (ainsi, tous les sous-programmes ANOVA n'accepteront pas que les volumes de données diffèrent selon les catégories).

Cet algorithme comporte les étapes suivantes:

Etape 1: Commencer par un tableau de données d'entrée à N lignes, chaque ligne correspondant à une situation (c'est-à-dire à une configuration source vidéo/distorsion donnée). Chaque ligne i comprend: le numéro de la source, le numéro de la distorsion, la note O_i de mesure VQM, le nombre de réponses N_i, la note subjective moyenne $S_{i\bullet}$ et la variance V_i des notes subjectives.

Etape 2: Transformer les notes subjectives $S_{i\bullet}$ en notes $\hat{S}_{i\bullet}$ de l'échelle commune de notation, comme on le décrit au § 2.1. La variance V_i des notes subjectives doit également être normalisée comme suit:

$$\hat{V}_i = \frac{V_i}{(worst - best)^2}$$

Il convient de noter que la transformation des notes subjectives et de leur variance est optionnelle. Celle-ci n'a pas d'incidence sur les données statistiques z définies ci-après mais peut modifier le processus d'ajustement VQM. Projeter ensuite les notes VQM O_i sur l'échelle commune de notation par l'intermédiaire d'une fonction d'ajustement, comme on l'indique dans le § 2.2 et le détaille dans l'Appendice 3. Le résultat de ce processus d'ajustement est un ensemble de notes VQM $\hat{O}_i$ figurant sur l'échelle commune de notation. Afficher les valeurs des coefficients d'ajustement utilisés ainsi que le domaine VQM sur lequel l'ajustement a été effectué (domaine de validité).

Etape 3: Pour chaque paire de situations distinctes d'indice i et j ($i \neq j$), utiliser un test z unidirectionnel pour attribuer une probabilité de signification à la différence entre la note VQM la plus grande et la note VQM la plus petite (respectivement $\hat{O}_i$ et $\hat{O}_j$). Cette signification est la probabilité que la note VQM la plus grande ait été attribuée à la situation dont la note subjective moyenne implicite réelle est la plus élevée. Cette note z est donnée par:

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

La probabilité de signification $p(z)$ de la note z est alors tout simplement égale à la fonction de distribution cumulative de z:

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

Etape 4: Tracer une courbe de dispersion de $p(z)$ (en ordonnée) en fonction de la note ΔVQM (en abscisse). Pour N situations, considérer chaque paire (i, j) avec $i > j$, enregistrer la différence VQM $\hat{O}_i - \hat{O}_j$ dans l'élément ΔVQM (d'indice k) du vecteur ΔVQM de longueur $N(N-1)/2$ et enregistrer la note z correspondante dans un vecteur Z de longueur $N(N-1)/2$ (pour l'élément de même indice k). On souhaite s'assurer que ΔVQM(k) est toujours positif, ce qui peut être obtenu par le biais de la définition de l'ordre a priori arbitraire des points i et j. Pour ce faire, si ΔVQM(k) est négatif, remplacer $Z(k)$ par $-Z(k)$ et ΔVQM(k) par $-$ΔVQM(k).

Etape 5: Considérer 19 segments (d'indice m) de ΔVQM, dont chacun occupe 1/10ème de l'intervalle total ΔVQM. Les segments se chevauchent de 50%. Associer ΔVQM_m au milieu de chaque segment et associer p_m à la valeur moyenne de $p(z)$ pour toutes les valeurs de z du segment m.

Etape 6: Tracer une courbe passant par les points (ΔVQM_m, p_m) pour obtenir une représentation graphique de p en fonction de ΔVQM. Noter que p peut être interprété comme étant la probabilité moyenne de signification.

Etape 7: Choisir un seuil de probabilité *p*. Tracer une ligne horizontale passant par l'ordonnée *p*: son intersection avec la courbe de l'Etape 6 détermine une valeur ΔVQM seuil, définie comme étant la probabilité. Pour une probabilité moyenne de signification égale ou supérieure à *p*, ΔVQM doit être supérieure à cette valeur seuil. La valeur de *p* généralement choisie est égale à 0,68, 0,75, 0,90 ou 0,95.

Ayant été déterminée pour la valeur *p* choisie, la valeur de ΔVQM peut être utilisée directement sur l'échelle commune de notation – démarche appropriée si l'on souhaite procéder au contre-étalonnage décrit à l'Etape 6. Il est également possible, à d'autres fins, de procéder à une projection inverse de cette valeur ΔVQM sur l'échelle de notation initiale, afin d'obtenir une valeur *R* du pouvoir de résolution sur l'échelle initiale en fonction de la note objective initiale *O*:

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

où *F* est la fonction d'ajustement définie au § 2.2. Concernant les fonctions logistiques décrites au § 2.2, l'inverse de la fonction logistique I est:

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

et l'inverse de la fonction logistique II est:

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

Lorsque $|\Delta\text{VQM}| << 1$, on peut définir pour $R(O)$ la valeur approchée suivante:

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

où $F'(O)$ est la dérivée de *F* par rapport à *O*. Cette approximation devrait être suffisante dans la plupart des cas.

NOTE 1 – Concernant les fonctions logistiques du § 2.2, la dérivée de la fonction logistique I est:

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

et la dérivée de la fonction logistique II est:

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

2.4 MÉTHODE DE MESURE 2: Calcul de l'erreur quadratique moyenne (RMSE, *root-mean-squared error*) associée à une méthode VQM

Si les données subjectives présentent à peu près une même variance sur l'échelle VQM, il peut être intéressant d'avoir une estimation de variance commune, ou pouvoir de résolution. Nous choisissons à titre d'exemple le RMSE. L'idée de base qui sous-tend le calcul de la valeur RMSE associée à VQM est de quantifier l'erreur quadratique moyenne (MSE, *mean square error*) existant entre les données objectives ajustées et les données subjectives correspondantes. La valeur RMSE associée

à une méthode VQM entre les données objectives ajustées $\hat{O}_i$ et les données subjectives normalisées $\hat{S}_{i\bullet}$ se calcule comme suit:

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

où:

N: nombre total de situations (égal au produit IJ, où J est le nombre de scènes et I est le nombre de conduits HRC)

D: degré de liberté intervenant pour l'ajustement de courbe entre données objectives et données subjectives mentionné au § 2.2.

2.5 Courbes de classification

Les erreurs de classification sont un moyen d'évaluer l'efficacité d'une méthode VQM. On parle d'erreur de classification lorsque le test subjectif et la méthode VQM conduisent à des conclusions différentes concernant une paire de points de données. Le présent paragraphe porte sur la signification des erreurs de classification, que l'on représente par les courbes des notes z subjectives en fonction des valeurs ΔVQM décrites dans le corps de ce texte. Dans la description qui suit, nous utilisons l'échelle commune de notation [0, 1] pour les notes subjectives et les notes objectives. «0» correspond à une absence de dégradation et «1» à une dégradation maximale.

Pour tout test subjectif, on peut fixer un seuil Δz qui permet de définir l'équivalence ou la distinction statistique entre deux points de données (A, B)[1]. Les résultats du test subjectif permettent ensuite de placer chaque paire de points de données (A, B) dans l'une des trois catégories suivantes:

$$\Delta z_{AB} < -\Delta z \quad \rightarrow A \text{ est supérieur à } B \quad \rightarrow Bs$$

$$-\Delta z \le \Delta z_{AB} \le \Delta z \quad \rightarrow A \text{ est équivalent à } B \quad \rightarrow Es$$

$$\Delta z < \Delta z_{AB} \quad \rightarrow A \text{ est inférieur à } B \quad \rightarrow Ws$$

Les abréviations utilisées pour ces trois catégories (*Bs*, *Es* et *Ws*) correspondent respectivement aux termes anglais *subjectively better* (subjectivement supérieur), *subjectively equivalent* (subjectivement équivalent) et *subjectively worse* (subjectivement inférieur).

Considérons à présent un seuil similaire Δo associé aux valeurs VQM:

$$\text{VQM}(A) - \text{VQM}(B) < -\Delta o \quad \rightarrow A \text{ est supérieur à } B \quad \rightarrow Bo$$

$$-\Delta o \le \text{VQM}(A) - \text{VQM}(B) \le \Delta o \quad \rightarrow A \text{ est équivalent à } B \quad \rightarrow Eo$$

$$\Delta o < \text{VQM}(A) - \text{VQM}(B) \quad \rightarrow A \text{ est inférieur à } B \quad \rightarrow Wo$$

Les abréviations utilisées pour ces trois catégories (*Bo*, *Eo* et *Wo*) correspondent respectivement aux termes anglais *objectively better* (objectivement supérieur), *objectively equivalent* (objectivement équivalent) et *objectively worse* (objectivement inférieur).

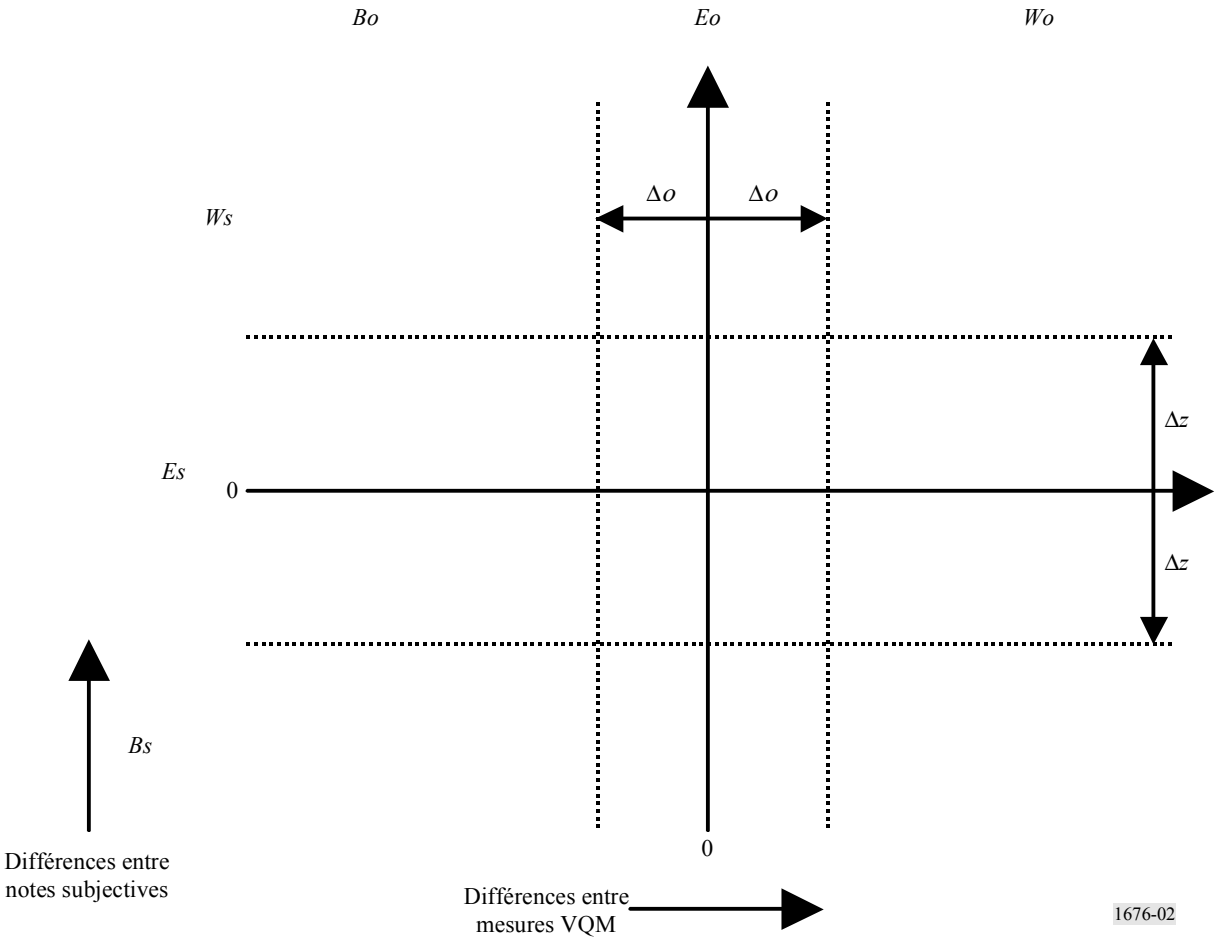
---

[1] Les points de données A et B correspondent en fait aux ensembles d'observations pour deux associations SRC/HRC. Comme on l'a vu dans le corps principal du texte, la quantité Δz_{AB} correspond à la différence des valeurs moyennes pour A et B ($\hat{S}_{A\bullet} - \hat{S}_{B\bullet}$) que divise l'écart type correspondant $\sqrt{(\hat{V}_A/N_A + \hat{V}_B/N_B)}$, où $\hat{V}_A$ est la variance des notes associées à la situation A et N_A est le nombre d'observations associées à la situation A, etc.

Chaque paire de données faisant l'objet d'une classification à trois catégories aux termes du test subjectif et d'une autre classification à trois catégories aux termes des mesures VQM, neuf types de résultat sont possibles. Ces neuf domaines sont séparés sur la Fig. 2 par des lignes pointillées dans l'espace à deux dimensions associé à l'axe de la différence entre notes subjectives et à l'axe de la différence entre mesures VQM.

FIGURE 2

Diagramme de classification schématique



Dans le Tableau ci-dessous, nous étiquetons chacune des neuf catégories de résultats en ayant pour optique de répondre à la question suivante: «Comment juge-t-on la classification à trois catégories fondée sur les mesures VQM au regard de la classification à trois catégories fondée sur le test subjectif?».

	Bs	***Es***	***Ws***
Wo	Classification erronée	Différentiation erronée	Décision correcte
Eo	Equivalence erronée	Décision correcte	Equivalence erronée
Bo	Décision correcte	Différentiation erronée	Classification erronée

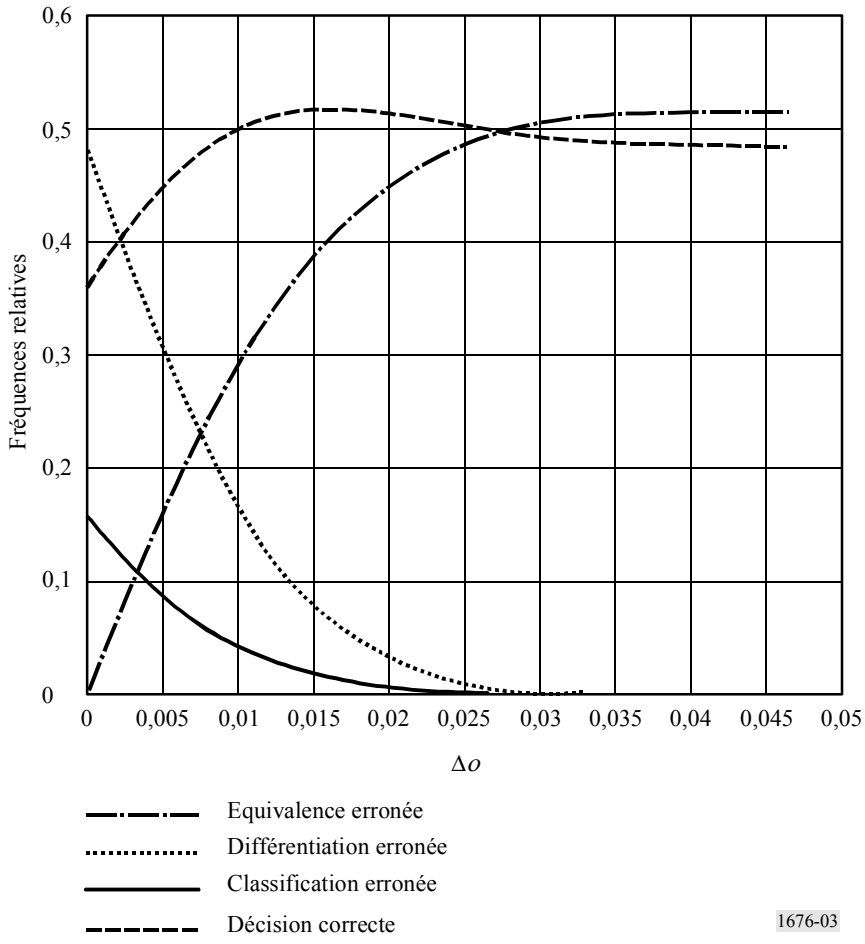
Il convient de noter que pour trois types de résultats, la classification VQM est identique à celle du test subjectif. On parle alors de «décision correcte». Les six catégories de résultats restants correspondent aux trois différents types d'erreurs que peut générer l'application d'une méthode VQM.

L'erreur de type «équivalence erronée» est sans doute la moins gênante. Elle se produit lorsque deux points de données sont jugés différents au vu du test subjectif alors qu'ils sont identiques au vu des mesures VQM. Une erreur de type «différentiation erronée» est généralement plus gênante. Elle apparaît lorsque, au vu du test subjectif, deux points de données sont identiques alors qu'ils sont différents au vu des mesures VQM. Une erreur de type «classification erronée» est généralement la plus pénalisante. Dans un tel cas, le test subjectif indique que *A* est supérieur à *B* alors que la méthode VQM indique le contraire.

Pour tout test subjectif et toute méthode VQM, on peut former toutes les paires distinctes possibles de points de données et compter le nombre de paires apparaissant dans chacune des quatre catégories de résultats: décision correcte, équivalence erronée, différentiation erronée et classification erronée. Nous pouvons ensuite normaliser ces nombres de paires par le nombre total de paires distinctes et reporter les fréquences relatives ainsi obtenues pour ces quatre catégories de résultats. Ces résultats dépendront généralement de Δs et de Δo. On trouvera à titre d'exemple les résultats associés à une méthode VQM fictive sur la Fig. 3. Δz a été choisi pour donner un intervalle de confiance à 95% pour les classifications subjectives et Δo est le paramètre variable sur l'axe *x* de la Figure.

FIGURE 3

Exemples de courbes représentant la fréquence des erreurs de classification



Il convient de noter qu'à mesure que Δo s'accroît, de plus en plus de paires de points seront considérées comme équivalentes après application de la méthode VQM. Cela réduit d'autant l'occurrence des différentiations erronées et des classifications erronées, mais accroît l'occurrence des équivalences erronées. Lorsque Δo tend vers 0,05, le taux d'équivalences erronées tend vers 0,52. La

méthode VQM indique alors que toutes les paires de points sont équivalentes, à tort dans 52% des cas et à juste titre pour 48% des occurrences. Ce résultat est cohérent avec le fait que pour ce test, 48% des paires de points de données ont été déclarées équivalentes au vu du test subjectif. Un graphique comme celui-ci pourrait être utilisé pour choisir une valeur appropriée de Δo. Par exemple, on pourrait choisir Δo de manière à maximiser la probabilité de prendre des décisions correctes ou en vue de minimiser une certaine somme pondérée de fréquences relatives d'erreurs.

Dans le code ayant servi à générer la Fig. 3 (une partie du code MATLAB de l'Appendice 2), le seuil utilisé pour le test subjectif est noté subj_th. Le seuil utilisé pour la mesure ΔVQM, vqm_th, est le paramètre variable. L'algorithme permet d'afficher en fonction de vqm_th les fréquences d'occurrence pour les trois types d'erreurs et pour l'absence d'erreur. Une valeur optimale de vqm_th pourrait être celle qui maximise la fréquence d'occurrence de l'absence d'erreur ou celle qui minimise une somme pondérée d'erreurs. Généralement, les équivalences erronées seront considérées comme les erreurs les moins gênantes, les différentiations erronées comme des erreurs plus gênantes et les classifications erronées comme les erreurs les plus pénalisantes.

NOTE 1 – L'utilisation de neuf catégories de résultats et d'un quadrillage divisant l'espace (ΔVQM, notes Z subjectives) en trois bandes verticales et trois bandes horizontales correspond à la description la plus naturelle de la présente analyse. Cette représentation suppose que ΔVQM peut être positif ou négatif, alors que l'algorithme utilise en fait la valeur absolue de ΔVQM (et remplace Z par $-Z$ pour tous les points dont l'abscisse ΔVQM est négative). Cela ne modifie aucunement les résultats mathématiques, mais conduit à une représentation plus naturelle faisant à présent appel à six catégories de résultats et à un espace à 2 bandes verticales et 3 bandes horizontales. Deux catégories de résultats associés à des décisions correctes (A supérieur à B ou A inférieur à B) n'en forment plus qu'une. On trouve alors encore deux catégories d'équivalences erronées, mais qu'une seule catégorie de différentiations erronées et une seule catégorie de classifications erronées.

3 Contre-étalonnage de deux méthodes VQM

On parvient à comparer deux méthodes VQM en utilisant la transformation (projection) vers l'échelle commune de notation décrite aux § 2.1 et 2.2[2]. Une fois déterminées les transformations permettant de projeter les mesures associées aux deux méthodes VQM (appelons-les VQM1 et VQM2) considérées sur l'échelle commune de notation (par le biais d'un ensemble convenu de données subjectives), la transformation pour passer de VQM1 à VQM2 consiste simplement à appliquer la transformation directe de VQM1 à l'échelle commune de notation puis la transformation indirecte de cette échelle commune à VQM2. Les modèles à comparer doivent être rapportés à un même ensemble de données. Lorsque les domaines ou les intervalles de données ne coïncident pas, on doit considérer que le contre-étalonnage n'est pas défini. La présente Recommandation ne spécifie pas un ensemble particulier de données communes.

---

[2] AVERTISSEMENT: On veillera à ne pas interpréter abusivement les résultats du contre-étalonnage – ainsi, contre-étalonner deux méthodes VQM ne signifie pas que l'une peut être remplacée par l'autre sans erreur. Une des raisons de cette limitation est que la méthode d'étalonnage actuelle dépend des données subjectives particulières qui définissent l'échelle commune de notation. On pourrait soutenir qu'aucune donnée subjective n'est nécessaire pour un contre-étalonnage et que l'on pourrait comparer deux méthodes VQM directement à partir des résultats de sortie associés à un ensemble particulier de données d'entrée (les paires source/circuit vidéo de référence testées). Cependant, quel que soit le jeu de données d'entrée choisi pour le contre-étalonnage, les résultats donnés par une méthode VQM pourraient être différents pour d'autres séquences vidéo. Plus fondamentalement, même pour l'ensemble de données sélectionné, il y a vraisemblablement quatre entrées (1, 2, 3, 4) de telle sorte que les notes VQM sont modifiées dans le même sens lorsque l'on passe de 1 à 2, mais en sens opposé lorsque l'on passe de 3 à 4. Ce type de comportement est ce qui fait qu'une méthode VQM est meilleure qu'une autre et aucune méthode de contre-étalonnage ne peut le refléter.

Appendice 1
à l'Annexe 1

Application de la présente Recommandation à l'évaluation et à la validation des méthodes VQM proposées

(Appendice informatif)

1 Eléments d'une présentation exhaustive d'une méthode VQM

Chaque méthode VQM à considérer doit être validée de façon indépendante et présentée de façon exhaustive de telle sorte qu'elle puisse être mise en œuvre sans difficulté par une personne disposant des connaissances adéquates. La description des méthodes VQM nouvellement proposées devrait comprendre les trois jeux de données suivants:

– les vecteurs test pour vérifier la mise en œuvre de la méthode VQM, notamment les données vidéo d'entrée et les résultats VQM de sortie;

– les données de validation/de précision, y compris les évaluations subjectives et les résultats de sortie du modèle VQM (la plage de qualité doit être suffisamment large pour être représentative des données vidéo transmises types);

– les données relatives à d'autres méthodes d'évaluation telles que le coefficient de corrélation linéaire de Pearson entre notes objectives et notes subjectives, le coefficient de corrélation de rang de Spearman d'ordre de corrélation entre notes objectives et notes subjectives et la fraction de points éloignés.

Enfin, le domaine et les limites d'application, la précision ainsi que le modèle de contre-étalonnage associés à cette méthode devraient être décrits conformément aux indications données dans les paragraphes suivants.

2 Domaine/limites d'application d'une méthode VQM

Le domaine d'application d'une méthode VQM peut faire référence aux éléments suivants (il s'agit là d'une liste donnée à titre illustratif, sans visée normative ou exhaustive):

– le type de contenu de la scène (le signal), par exemple un mouvement rapide ou lent, une image couleur ou noir et blanc, un balayage progressif ou à entrelacement;

– le type et l'intensité des artéfacts (le bruit) induits par les techniques de codage et les débits binaires (par exemple le flou, le tuilage);

– les conditions d'observation (notamment la distance d'observation, l'éclairement ambiant et les paramètres d'affichage tels que le paramètre gamma, la luminosité ou les types de phosphore).

Chaque méthode VQM devrait être évaluée qualitativement du point de vue du type de contenu de la scène, du type et de l'intensité des artéfacts et des conditions d'observation nécessaires pour une application efficace de la méthode. Il est important d'énumérer les zones connues de difficultés (distorsions vidéo par perte d'images par exemple) dont la connaissance n'est pas évidente, même si la partie consacrée au domaine/limites d'application de la méthode ne doit pas prétendre à l'exhaustivité.

Quatre tableaux devraient être fournis pour décrire le domaine et les limites d'application de la méthode VQM. Les trois premiers devraient dresser la liste de toutes les distorsions (sources HRC) pour l'ensemble de données du groupe d'experts sur la qualité vidéo (VQEG, *Video Quality Experts Group*), voire éventuellement d'autres distorsions. Il s'agit des tableaux suivants:

– un tableau de facteurs de test, de techniques de codage et d'applications pour lesquels la méthode VQM a satisfait au degré de précision spécifié;

– un tableau de facteurs de test, de techniques de codage et d'applications pour lesquels la méthode VQM a été testée mais n'a *pas* satisfait au degré de précision spécifié dans le § 2;

– un tableau de facteurs de test, de techniques de codage et d'applications connus (y compris toutes celles utilisées par le groupe VQEG) pour lesquels la méthode VQM n'a pas été testée ou n'est pas recommandée.

Il faudrait en outre disposer d'un tableau des séquences de test utilisées pour déterminer les facteurs de test, les techniques de codage et les applications pour lesquels la méthode VQM a satisfait au degré de précision spécifié dans le § 2.

On trouvera ci-dessous les Tableaux donnés à titre d'exemple.

TABLEAU 1

a) Facteurs de test, techniques de codage et applications pour lesquels la méthode VQM à considérer a satisfait au degré de précision spécifié

Débit binaire	Résolution	Méthode	Observations
2 Mbit/s	Résolution en 3/4	mp@ml	Diminution de la résolution horizontale uniquement
2 Mbit/s	Résolution en 3/4	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	Composite NTSC et/ou PAL
6 Mbit/s		mp@ml	
8 Mbit/s		mp@ml	Composite NTSC et/ou PAL
8 & 4,5 Mbit/s		mp@ml	Deux codecs concaténés
19/PAL(NTSC)- 19/PAL(NTSC)- 12 Mbit/s		422p@ml	PAL ou NTSC 3 générations
50-50-… –50 Mbit/s		422p@ml	7ème génération avec décalage par rapport à l'image I
19-19-12 Mbit/s		422p@ml	3ème génération
Sans objet		Sans objet	Betacam multigénération avec suppression (4 ou 5, signal composite/en composantes)

TABLEAU 1 (*fin*)

b) Facteurs de test, techniques de codage et applications pour lesquels la méthode VQM n'a pas satisfait au degré de précision spécifié

Débit binaire	Résolution	Méthode	Observations
4,5 Mbit/s		mp@ml	Avec erreurs
3 Mbit/s		mp@ml	Avec erreurs

Ces Tableaux utilisant toutes les données VQEG, le Tableau 1c est sans objet.

d) Séquences de test utilisées pour déterminer les facteurs de test, les techniques de codage et les applications pour lesquels la méthode VQM a satisfait au degré de précision spécifié

Séquence	Caractéristiques
Balloon-pops	Film, couleurs saturées, mouvement
NewYork 2	Effet de masque, mouvement
Mobile & Calendar	Disponible en deux formats, couleur, mouvement
Betes_pas_betes	Couleur, synthétique, mouvement, scène interrompue
Le_point	En couleur, transparence, mouvement dans toutes les directions
Autumn_leaves	En couleur, paysage, zoom, mouvement en cascade
Football	En couleur, mouvement
Sailboat	Image presque fixe
Susie	Couleur peau
Tempête	Couleur, mouvement

Appendice 2 à l'Annexe 1

Code source MATLAB

(Appendice informatif)

On trouvera ci-après le sous-programme MATLAB intitulé vqm_accuracy.m. Cette version de code projette les données subjectives sur l'intervalle [0, 1], applique un ajustement polynomial pour permettre une mise en correspondance entre données objectives et données subjectives normalisées, calcule toutes les grandeurs de mesure et affiche les fréquences d'occurrence VQM d'«équivalences erronées», de «différentiations erronées», de «classifications erronées» et de «décisions correctes». La version 5.3.1 de MATLAB (1999) et ses outils logiciels de calculs statistiques et d'optimisation, disponibles séparément, sont suffisants. L'élaboration d'un algorithme ne faisant appel à aucun de ces deux outils logiciels est également possible. Le code présenté ci-après est donné à titre illustratif et ne fait pas appel à toutes les options et fonctions d'ajustement possibles.

Utilisation: à partir de l'invite de commande Matlab et pour la méthode VQM r0, entrer:

>load r0.dat

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

Pour la méthode VQM r2, entrer:

>load r2.dat

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

r0.dat et r2.dat sont ici des fichiers textes qui contiennent un sous-ensemble de données de 525 lignes VQEG. Chaque ligne du fichier correspond à une situation et comprend le numéro de source SRC, le numéro de circuit HRC, la note VQM, le nombre d'observations, la note subjective moyenne et la variance associée à cette note subjective moyenne. Une fois les fichiers r0 et r2.dat chargés, l'exécutable vqm_accuracy peut à nouveau être lancé pour r0 ou pour r2.

Concernant le premier argument de vqm_accuracy, r0 correspond au modèle PSNR associé à TR A3 et r2 correspond au modèle PQR associé à TR A4. Le deuxième argument a pour valeur 1 si la valeur de la mesure objective augmente lorsque la qualité d'image diminue, et –1 dans le cas contraire. Le troisième et le quatrième argument correspondent aux notes maximale et minimale sur l'échelle de notation subjective initiale. Le dernier argument correspond à l'ordre du polynôme utilisé pour la courbe d'ajustement des mesures VQM.

Code source:

```
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
```

Appendice 3
à l'Annexe 1

Ajustement des données VQM à une échelle commune de mesures

(Appendice informatif)

Comme on l'a vu au § 2.2 du corps principal du document, les données VQM objectives (O_i) sont projetées sur un nouveau domaine et deviennent $\hat{O}_i = F(O_i)$. Ce domaine est obtenu par ajustement des données O_i aux données subjectives normalisées ($\hat{S}_{i\bullet}$) à l'aide d'une famille de fonctions F (avec leurs paramètres d'ajustement) ayant les propriétés de monotonie et de mise en correspondance des intervalles décrites au § 5.2. On trouvera ci-après un choix de trois types de fonctions F ainsi que des notes sur l'ajustement de données par le biais de ces fonctions.

1 Polynôme d'ordre *M*

Un polynôme d'ajustement à un ensemble de points de données n'est pas forcément une fonction monotone. L'outil d'optimisation MATLAB dispose d'une fonction lsqlin qui assure la monotonie sur le domaine de variation des données. Toutefois, la monotonie sur le domaine de données existant ne garantit pas la monotonie sur l'intégralité du domaine théorique (par exemple de 0 à l'infini).

2 Fonction logistique I

L'ajustement des données VQM objectives (O_i) aux données subjectives normalisées ($\hat{S}_{i\bullet}$) peut être fait à l'aide de la fonction logistique:

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

où a, b, c, d et e sont les paramètres d'ajustement. La fonction d'ajustement doit être obtenue à partir de la méthode non linéaire des moindres carrées (voir les notes MATLAB dans le Rapport final VQEG[3]). La partie de la fonction à utiliser est la partie monotone pour $O > -d$ (d'où la contrainte ($d > -\min(O)$) et la forme de la courbe en s appropriée pour l'ajustement des données est assurée par le respect de la contrainte $e > 1$.

Dans certains cas, au moins asymptotiquement, la note parfaite du modèle objectif sur l'échelle initiale peut être associée à la valeur zéro (c'est-à-dire la meilleure note sur l'échelle subjective) et la note objective possible la plus mauvaise sur l'échelle initiale peut être mise en correspondance avec la note subjective la plus mauvaise (c'est-à-dire l'unité sur l'échelle commune de notation). Considérons par exemple le cas suivant: zéro est la meilleure note objective et l'infini correspond à la note objective la plus mauvaise. On doit ainsi faire correspondre, lors du changement d'échelle, zéro à zéro et l'infini à 1, ce qui conduit à: $a = 1$ et $b = -(1 + cd^e)$. On a donc pour F:

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

L'ajustement portera sur c, d et e, sous réserve du respect des conditions $d > 0$ et $e > 0$.

---

[3] L'ajustement par courbe non linéaire des moindres carrées avec contraintes peut être réalisé par la fonction MATLAB lsqcurvefit.

3 Fonction logistique II

Le passage des données VQM objectives (O_i) aux données subjectives normalisées ($\hat{S}_{i\bullet}$) peut également se faire en utilisant la fonction logistique:

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

où a, b, c et d sont les paramètres d'ajustement et $c > 0$ (ce qui est garanti en définissant $c = |C|$ pour C réel). Comme dans le cas de la fonction logistique I, la fonction d'ajustement doit être obtenue à partir de la méthode non linéaire des moindres carrées[4].

On pourrait utiliser cette optimisation dans le cas noté § 2: zéro est la meilleure note objective et l'infini correspond à la note objective la plus mauvaise. Ainsi, zéro correspond à zéro et l'infini correspond à 1, ce qui conduit à: $a = -\exp[-cd]$ et $b = -a \exp[cd]$. On a donc pour F:

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

La fonction logistique II est également utile dans le cas suivant (ce qui pourrait arriver si O_i était exprimée en coordonnées logarithmiques, par exemple en décibels): la meilleure note objective est l'infini et la note objective la plus mauvaise est moins l'infini. Dans ce cas, l'infini doit être associé à zéro et moins l'infini à 1. On a alors $b = 0$, $a = 1$ et:

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

---

[4] On peut lire à la page 31 du Rapport final VQEG que les valeurs initiales des paramètres ont été choisies comme suit: a = note subjective maximale, b = note subjective minimale, $c = 1$ et d = note objective moyenne.