ITUPublications
Recommandations

Union internationale des télécommunications
Secteur des radiocommunications

Recommandation UIT-R BT.500-15

(05/2023)

Série BT: Service de radiodiffusion télévisuelle

Méthodologies d'évaluation subjective de la qualité des images de télévision

Avant-propos

Le rôle du Secteur des radiocommunications est d'assurer l'utilisation rationnelle, équitable, efficace et économique du spectre radioélectrique par tous les services de radiocommunication, y compris les services par satellite, et de procéder à des études pour toutes les gammes de fréquences, à partir desquelles les Recommandations seront élaborées et adoptées.

Les fonctions réglementaires et politiques du Secteur des radiocommunications sont remplies par les Conférences mondiales et régionales des radiocommunications et par les Assemblées des radiocommunications assistées par les Commissions d'études.

Politique en matière de droits de propriété intellectuelle (IPR)

La politique de l'UIT-R en matière de droits de propriété intellectuelle est décrite dans la «Politique commune de l'UIT-T, l'UIT-R, l'ISO et la CEI en matière de brevets», dont il est question dans la Résolution UIT-R 1. Les formulaires que les titulaires de brevets doivent utiliser pour soumettre les déclarations de brevet et d'octroi de licence sont accessibles à l'adresse http://www.itu.int/ITU-R/go/patents/fr, où l'on trouvera également les Lignes directrices pour la mise en œuvre de la politique commune en matière de brevets de l'UIT-T, l'UIT-R, l'ISO et la CEI et la base de données en matière de brevets de l'UIT-R.

Séries des Recommandations UIT-R

(Également disponible en ligne: https://www.itu.int/pub/R-REC/fr)

Séries	Titre
BO	Diffusion par satellite
BR	Enregistrement pour la production, l'archivage et la diffusion; films pour la télévision
BS	Service de radiodiffusion sonore
BT	**Service de radiodiffusion télévisuelle**
F	Service fixe
M	Services mobile, de radiorepérage et d'amateur y compris les services par satellite associés
P	Propagation des ondes radioélectriques
RA	Radio astronomie
RS	Systèmes de télédétection
S	Service fixe par satellite
SA	Applications spatiales et météorologie
SF	Partage des fréquences et coordination entre les systèmes du service fixe par satellite et du service fixe
SM	Gestion du spectre
SNG	Reportage d'actualités par satellite
TF	Émissions de fréquences étalon et de signaux horaires
V	Vocabulaire et sujets associés

***Note**: Cette Recommandation UIT-R a été approuvée en anglais aux termes de la procédure détaillée dans la Résolution UIT-R 1.*

Publication électronique
Genève, 2024

RECOMMANDATION UIT-R BT.500-15

Méthodologies d'évaluation subjective de la qualité des images de télévision[1]

(Question UIT-R 102-4/6)

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

Domaine d'application

La présente Recommandation contient des méthodes d'évaluation subjective de la qualité des images, notamment les méthodes générales d'essai, les échelles de notation utilisées lors des évaluations et les conditions d'observation recommandées pour effectuer les évaluations. Elle est composée de trois parties:

– La Partie 1 décrit les exigences générales pour procéder à l'évaluation subjective des images de télévision et des orientations concernant les circonstances d'utilisation de telle ou telle méthode.

– La Partie 2 décrit les différentes méthodes d'évaluation recommandées qu'il est possible d'utiliser lorsqu'on effectue des évaluations subjectives de la qualité des images.

– La Partie 3 décrit des méthodes propres à certains formats d'images ou certaines applications sur la base des spécifications données dans les Parties 1 et 2.

Mots clés

Évaluation subjective, évaluation des images

L'Assemblée des radiocommunications de l'UIT,

considérant

a) qu'une grande quantité d'informations a été recueillie sur les méthodes utilisées dans divers laboratoires pour l'évaluation de la qualité de l'image;

b) que l'examen de ces méthodes montre qu'il existe un large accord entre différents laboratoires sur un certain nombre d'aspects des méthodes d'évaluation subjective;

c) qu'il est important d'adopter des méthodes d'évaluation normalisées pour l'échange d'informations entre les divers laboratoires;

d) que les évaluations de la qualité ou de la dégradation de l'image faites en exploitation normale ou spéciale par certains techniciens chargés du contrôle, en utilisant des échelles à cinq notes, peuvent également s'inspirer de certains aspects des méthodologies recommandées pour les essais en laboratoire;

e) que l'introduction en permanence de nouveaux signaux de télévision, de nouveaux modes de traitement des signaux et de nouveaux services de télévision évolués peut nécessiter des méthodes différentes pour effectuer des évaluations subjectives des images;

f) que l'introduction de tels traitements, signaux et services augmentera la probabilité selon laquelle la qualité de chaque section de la chaîne du signal dépendra de plus en plus des opérations effectuées en amont,

[1] La présente Recommandation doit être portée à l'attention de la Commission d'études 12 de l'UIT-T.

recommande

1 que les méthodologies générales d'essai, les échelles et les conditions d'observation pour l'évaluation de la qualité des images décrites dans la Partie 1 soient utilisées pour les expériences de laboratoire et aussi pour les évaluations en exploitation chaque fois que cela est possible;

2 que, malgré l'existence d'autres méthodologies et la mise au point de nouvelles méthodologies, celles décrites dans la Partie 2 de la présente Recommandation soient utilisées lorsqu'il y a lieu;

3 que les méthodologies générales d'essai, les échelles de notation et les conditions d'observation pour l'évaluation de la qualité des images d'un système ou d'une application d'images donné décrites dans la Partie 3 soient utilisées pour les expériences en laboratoire et, chaque fois que possible, pour les évaluations opérationnelles;

4 que, pour faciliter les échanges d'informations entre différents laboratoires, les exigences de la méthodologie d'essai choisie soient respectées comme indiquées dans la Partie 2;

5 que, pour faciliter les échanges d'informations entre différents laboratoires, les données recueillies soient traitées conformément aux techniques statistiques décrites à l'Annexe 2 de la Partie 1;

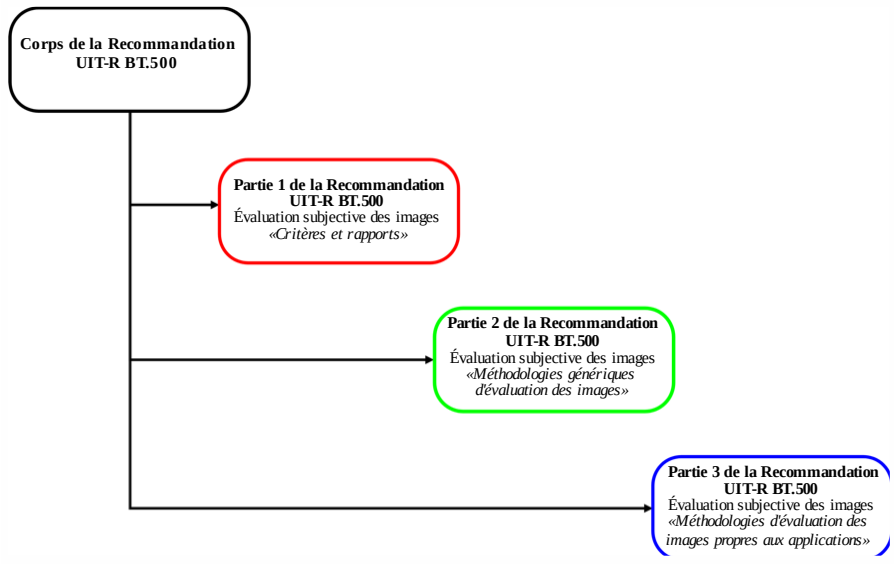
6 qu'étant donné qu'il est important de définir la base des évaluations d'image subjectives, tous les rapports d'essai donnent la description la plus complète possible des configurations d'essai, du matériel d'essai, des observateurs et des méthodes.

Notes concernant la structure et l'utilisation de la présente Recommandation (pour information)

La Recommandation UIT-R BT.500 est composée de trois Parties semi-autonomes regroupées dans la présente Recommandation, comme le montre la Fig. 1.

FIGURE 1

Structure de la Recommandation UIT-R BT.500



BT.0500-01

Il est recommandé aux laboratoires qui souhaitent effectuer des évaluations subjectives des images de consulter le *recommande* ci-dessus, puis d'utiliser les critères présentés dans la Partie 1 afin de comprendre quelle est la méthodologie la plus appropriée pour leurs procédures d'évaluation. La

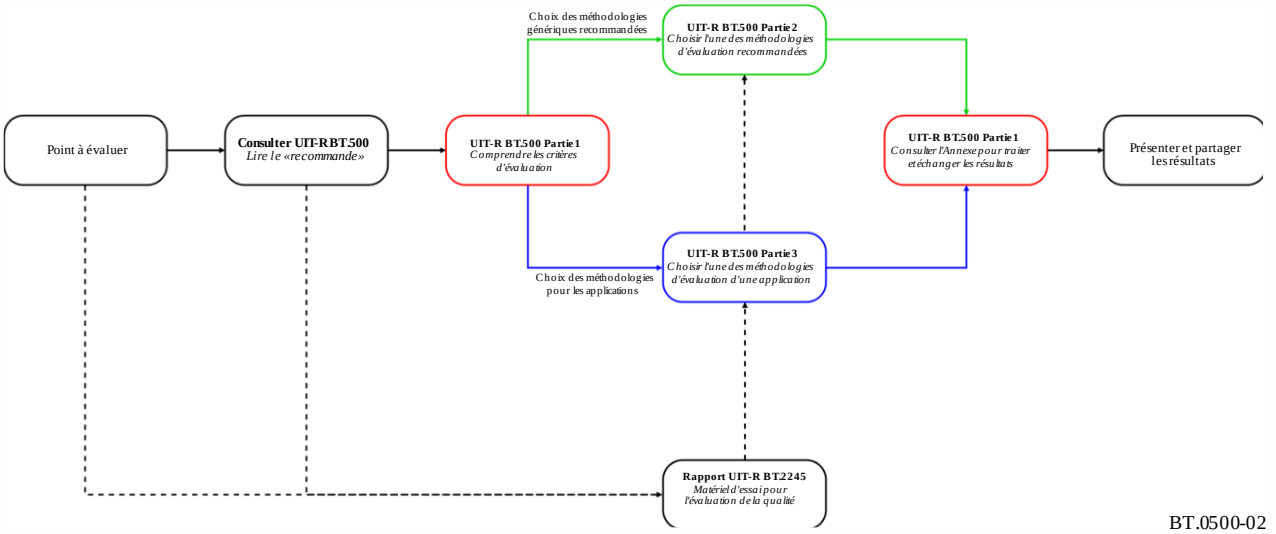
Partie 2 donne un aperçu de plusieurs méthodologies d'évaluation subjective des images recommandées pouvant être utilisées. La Partie 3 donne des informations sur d'autres méthodologies d'évaluation propres aux applications susceptibles de faciliter l'élaboration de procédures d'évaluation subjective des images connexes.

Conseils sur les modalités d'utilisation de la Recommandation UIT-R BT.500

La Figure 2 montre une possible manière de procéder pour utiliser la Recommandation UIT-R BT.500.

FIGURE 2

Utiliser la Recommandation UIT-R BT.500



Motif

Grâce à la structure organisée en plusieurs parties de la présente version de la Recommandation UIT-R BT.500, il est possible d'ajouter de nouvelles méthodologies d'évaluation subjective des images ou de réviser les méthodologies existantes sans qu'il soit nécessaire d'avoir de nouvelles Recommandations qui reprennent des informations dans de multiples documents ou de publier des révisions de parties qui n'ont pas besoin d'être modifiées.

Autres Recommandations portant sur l'évaluation des images

Les Recommandations ci-après portent sur la mesure objective de la qualité des images et peuvent fournir d'autres méthodologies d'évaluation des images pour les applications qui utilisent certains critères d'évaluation de la Recommandation UIT-R BT.500.

Recommandation UIT-R BT.1683	Techniques de mesure objective de la qualité vidéo perceptuelle pour la télédiffusion numérique à définition normale en présence d'une image de référence complète
Recommandation UIT-R BT.1866	Techniques de mesure objective de la qualité vidéo perceptuelle pour les applications de radiodiffusion utilisant la télévision basse définition en présence d'un signal de référence complet
Recommandation UIT-R BT.1867	Techniques de mesure objective de la qualité vidéo perceptuelle pour les applications de télédiffusion utilisant la télévision basse définition en présence d'un signal de référence à largeur de bande réduite

Recommandation UIT-R BT.1885	Techniques de mesure objective de la qualité vidéo perçue pour la télédiffusion numérique à définition normale en présence d'une largeur de bande réduite
Recommandation UIT-R BT.1907	Techniques de mesure objective de la qualité vidéo perçue pour les applications de radiodiffusion utilisant la télévision haute définition en présence d'un signal de référence complet
Recommandation UIT-R BT.1908	Techniques de mesure objective de la qualité vidéo pour les applications de radiodiffusion utilisant la télévision haute définition en présence d'un signal de référence réduit

PARTIE 1

Aperçu des exigences concernant l'évaluation subjective des images

1 Introduction

Les méthodes d'évaluation subjective des images servent à définir la qualité des systèmes de télévision au moyen de mesures qui tiennent compte avec précision des réactions de ceux qui observeront les systèmes à l'essai. On sait bien qu'à cet égard les méthodes objectives ne peuvent rendre exactement compte de la qualité d'un système et qu'il faut donc les compléter par des mesures subjectives.

On considère en général deux catégories d'évaluations subjectives. En premier lieu, celles qui établissent la qualité d'un système dans les meilleures conditions, appelées généralement évaluations de la qualité. En second lieu, celles qui établissent la faculté qu'ont les systèmes de conserver leur qualité dans des conditions non idéales de transmission ou d'émission, appelées généralement évaluations des dégradations.

Pour procéder aux évaluations subjectives les plus appropriées, il faut d'abord choisir, parmi les différentes options disponibles, la méthodologie qui convient le mieux aux conditions et aux objectifs spécifiques de l'évaluation des images requise.

Pour faciliter ce choix, les caractéristiques générales présentées dans le § 2 devraient être examinées afin de comprendre quelles sont les solutions les mieux appropriés pour le problème ou la procédure que l'on évalue.

Une fois ces options comprises, on trouvera dans le § 3 de la Partie 1 un aperçu des méthodologies d'évaluation des images recommandées, qui peut être utilisé pour faciliter le choix de la méthodologie la mieux adaptée au problème ou à la procédure que l'on évalue, compte tenu du type d'observateur utilisé et des conditions dans l'environnement d'évaluation.

Le choix de la méthodologie la plus appropriée dépend cependant des objectifs que le système à l'essai vise à atteindre. Les procédures complètes d'évaluation d'applications spécifiques font donc l'objet de la Partie 2 et d'autres Recommandations UIT-R.

2 Caractéristiques communes d'évaluation

On trouvera dans ci-après la description des conditions générales d'observation pour les évaluations subjectives. Les méthodologies connexes indiquent les conditions d'observation spécifiques pour des systèmes particuliers.

NOTE − Lors de l'évaluation subjective d'images à grande plage dynamique, il est conseillé de consulter d'autres documents cités en référence, lorsqu'ils sont disponibles, dans la section appropriée[2].

2.1 Conditions générales d'observation

L'environnement d'observation en laboratoire fournit des conditions extrêmes pour le contrôle des systèmes. Le § 2.1.1 spécifie les conditions générales d'observation pour les évaluations subjectives en laboratoire.

---

[2] La présente Recommandation sera révisée en vue d'y intégrer des orientations supplémentaires en fonction des avancées concernant les travaux et l'expérience acquise dans le domaine des images à grande plage dynamique.

L'environnement d'observation dans les domiciles fournit un moyen pour évaluer la qualité à l'extrémité «utilisateur» de la chaîne télévisuelle. Les conditions générales d'observation décrites au § 2.1.2 reproduisent l'environnement existant dans un domicile. Ces paramètres ont été choisis de manière à définir un environnement un peu plus critique que les conditions typiques d'observation dans les domiciles.

2.1.1 Conditions générales d'observation pour les évaluations subjectives en laboratoire

Les conditions d'observation doivent être les suivantes pour les évaluations:

a)	Éclairement de la salle:	faible
b)	Chromaticité de l'arrière-plan:	D_{65}
c)	Luminance de crête[3]:	70-250 cd/m^2 (Voir § 2.1.6.5)
d)	Rapport de contraste à l'écran:	≤ 0,02 (Voir § 2.1.6.4)
e)	Rapport luminance de l'arrière-plan, derrière l'écran d'affichage de l'image/luminance de crête de l'image:	≈ 0,15

2.1.2 Conditions générales d'observation pour les évaluations subjectives dans les domiciles

a)	Éclairement de l'écran lié à l'environnement (la lumière incidente provenant de l'environnement et qui arrive sur l'écran doit être mesurée perpendiculairement à l'écran):	200 lux
b)	Luminance de crête:	70-500 cd/m^2 (Voir § 2.1.6.4)
c)	Rapport luminance de l'écran inactif/rapport de contraste à l'écran:	≤ 0,02 (Voir § 2.1.6.4)

2.1.3 Distance d'observation

La distance d'observation est fonction de la taille de l'écran et peut être sélectionnée selon deux critères distincts: la distance d'observation préférée (PVD, *preferred viewing distance*) ou la distance d'observation nominale (DVD, *design viewing distance*). Le choix de l'une ou de l'autre dépendra de l'objet de l'étude.

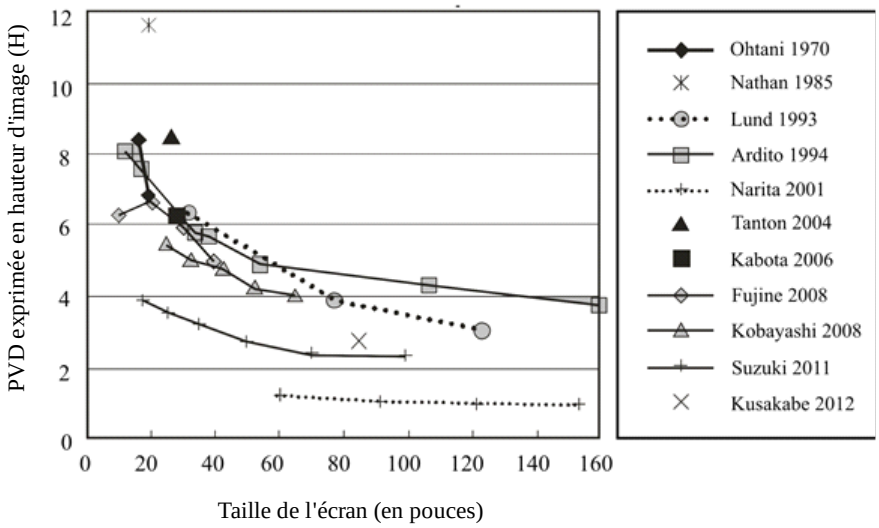
2.1.3.1 Distance d'observation préférée

La distance d'observation préférée (PVD) est fondée sur les préférences du téléspectateur, qui ont été déterminées empiriquement. La PVD (fonction des dimensions de l'écran) est indiquée dans la Fig. 1-1, qui contient plusieurs ensembles de données recueillies auprès des sources disponibles. Ces données peuvent servir de référence pour la conception d'un essai d'évaluation subjective.

[3] La luminance de crête doit être ajustée en fonction de l'éclairement de la salle.

FIGURE 1-1

Distance d'observation préférée en fonction des dimensions de l'écran



BT.0500-01-1

2.1.3.2 Distance d'observation nominale

Pour un système numérique, la distance d'observation nominale (DVD), ou distance d'observation optimale, est la distance à laquelle deux pixels adjacents forment un angle d'une minute d'arc depuis l'œil de l'observateur; et l'angle d'observation horizontal optimal correspond à la distance d'observation optimale d'une image.

Le Tableau 1-1 présente les distances d'observation optimales (et les angles d'observation horizontaux optimaux), exprimées en multiples de la hauteur de l'image, pour plusieurs résolutions d'image.

TABLEAU 1-1

Angle d'observation horizontal optimal, distance d'observation optimale exprimée en multiples de la hauteur d'image (H)

Résolution	Référence	Format d'image	Format de pixel	Angle d'observation horizontal optimal	Distance d'observation optimale
720 × 483	UIT-R BT.601	4:3	0,89	11°	7 *H*
640 × 480	VGA	4:3	1	11°	7 *H*
720 × 576	UIT-R BT.601	4:3	1,07	13°	6 *H*
1 024 × 768	XGA	4:3	1	17°	4.5 *H*
1 280 × 720	UIT-R BT.1543 et BT.1874	16:9	1	21°	4.8 *H*
1 400 × 1 050	SXGA+	4:3	1	23°	3,3 *H*
1 920 × 1 080	UIT-R BT.709	16:9	1	31°	3,2 *H*
3 840 × 2 160	UIT-R BT.2020	16:9	1	58°	1,6 *H*
7 680 × 4 320	UIT-R BT.2020	16:9	1	96°	0,8 *H*

Note: Lorsque l'évaluation de l'image porte sur la résolution, il convient d'utiliser la valeur de distance d'observation la plus faible pour les formats 7 680 × 4 320 et 3 840 × 2 160. Dans le cas contraire, on pourra choisir n'importe quelle distance dans l'intervalle indiqué (pour le format 3 840 × 2 160: de 1,6 fois à 3,2 fois la hauteur d'image; pour le format 7 680 × 4 320: de 0,8 fois à 3,2 fois la hauteur d'image).

2.1.4 Angle d'observation

L'angle d'observation maximal par rapport à la normale devrait être limité de manière à ce que les déviations des couleurs reproduites sur l'écran ne puissent être décelées par un observateur. Il convient également de prendre en considération l'angle d'observation horizontal optimal d'un système d'image à l'essai pour déterminer l'angle d'observation. Voir le § 1.8 du Rapport UIT-R BT.2129 pour en savoir plus.

2.1.5 Système de couleur dans l'environnement de la salle

La couleur de l'arrière-plan de l'écran devrait être identique à celle du point blanc de référence; les autres surfaces de la salle devraient être mates foncées. L'objectif est de réduire autant que possible la lumière parasite sur l'écran.

2.1.6 Caractéristiques générales de l'écran

La qualité subjective d'image obtenue sera différente selon les caractéristiques des écrans utilisées. C'est pourquoi il est vivement recommandé de vérifier au préalable les caractéristiques des écrans utilisées. La Recommandation UIT-R BT.1886 – Fonction de transfert électro-optique de référence pour les écrans plats utilisés pour la production en studio de TVHD, et le Rapport UIT-R BT.2129 – Besoins des utilisateurs en matière d'écran plat comme écran principal dans un environnement de production de programmes de TVHD, peuvent servir de références dans le cas de l'utilisation d'écrans plats professionnels aux fins d'une évaluation subjective.

Le Rapport UIT-R BT.2390 donne des informations sur les écrans utilisés en laboratoire et à domicile ainsi que sur les environnements d'observation pour l'évaluation des images à grande plage dynamique (HDR).

2.1.6.1 Traitement au niveau de l'écran

Il faudrait éviter de créer des artefacts lors du traitement au niveau de l'écran (mise à l'échelle de l'image, conversion de la fréquence d'image, amélioration des images, etc.), s'il est mis en œuvre. Le traitement HDR devrait être adapté au système HDR évalué ou utilisé pour l'évaluation. Pour des évaluations concernant l'environnement grand public ou la distribution, cela peut comprendre l'utilisation des métadonnées statiques ou dynamiques appropriées. Les informations complètes concernant ces métadonnées devraient figurer dans les notes des évaluations afin que d'autres laboratoires puissent répéter avec exactitude les évaluations.

Lorsque des écrans grand public sont utilisés pour l'évaluation subjective des images, il est important que toutes les options de traitement des images soient désactivées, sauf si l'évaluation porte sur l'incidence de ce traitement des images.

Lorsqu'on a accès à des images entrelacées, le rapport d'essai devrait indiquer si un désentrelaceur a été utilisé ou non. Il est préférable de ne pas utiliser de désentrelaceur si les signaux entrelacés peuvent être affichés sans.

2.1.6.2 Résolution de l'écran

La résolution des écrans à usage professionnel est généralement conforme aux normes requises pour les évaluations subjectives, en ce qui concerne la gamme de luminance dans laquelle ils fonctionnent.

On pourrait envisager de vérifier et de faire figurer dans un rapport la résolution maximale et minimale (au centre et dans les angles de l'écran) pour la valeur de luminance utilisée.

Si des téléviseurs écran plat grand public sont utilisés pour les évaluations subjectives, il est vivement recommandé de vérifier et de faire figurer dans un rapport la résolution maximale et la résolution minimale (au centre et dans les angles de l'écran) pour la valeur de luminance utilisée.

Actuellement le système le plus pratique dont disposent les responsables des évaluations subjectives pour vérifier le pouvoir de résolution des écrans ou des récepteurs de télévision grand public est un système à mire électronique avec balayage.

2.1.6.3 Réglage de l'écran

La luminosité et le contraste de l'écran devraient être réglés en fonction de l'éclairement lié à l'environnement au moyen des signaux PLUGE, conformément à la Recommandation UIT-R BT.814.

Pour l'évaluation des images à plage dynamique type (SDR), le niveau de contraste de l'écran devrait être mesuré conformément à la Recommandation UIT-R BT.815. Pour l'évaluation des images HDR, il convient de consulter le Rapport UIT-R BT.2390.

2.1.6.4 Contraste des écrans

Le contraste pourrait être fortement influencé par la luminance ambiante.

Il est rare que les écrans à usage professionnel mettent en œuvre des techniques permettant d'améliorer leur contraste dans des conditions de fort éclairement. Il est possible, par conséquent, que ces écrans ne respectent pas la norme de contraste requise lorsqu'ils fonctionnent dans de telles conditions.

Les écrans grand public utilisent en général des techniques pour améliorer le contraste dans des conditions de fort éclairement.

2.1.6.5 Luminosité de l'écran

Pour régler la luminosité d'un écran à cristaux liquides, il est préférable de modifier l'intensité de l'éclairage à contre-jour plutôt que d'utiliser le réglage du niveau du signal afin de conserver la précision binaire. Dans le cas d'autres technologies d'écran qui n'utilisent pas d'éclairage à contre-jour, il convient de régler le niveau du blanc par d'autres moyens que le réglage du niveau du signal. Il est à noter que, dans le cas des écrans plasma, le réglage de la luminosité se fait en modifiant l'intensité lumineuse et que, si l'on diminue la luminosité, la restitution des tons sera moins bonne.

2.1.6.6 Artéfacts de mouvement dus à l'écran

Les artéfacts de mouvement dus à la technologie d'écran particulière utilisée ne devraient pas apparaître sur l'écran. Par contre, les effets de mouvement inclus dans le signal d'entrée devraient apparaître à l'écran. Lorsqu'on utilise des écrans grand public, il est essentiel que TOUTES les options de traitement du mouvement soient désactivées.

2.1.6.7 Zones de sécurité des images produites au format écran large 16:9

Les zones de sécurité des images produites au format écran large 16:9 sont définies dans la Recommandation UIT-R BT.1848.

2.2 Signaux source

Le signal source fournit directement l'image de référence et le signal source pour le système à évaluer. Sa qualité doit être optimale pour la norme de télévision utilisée. Il est essentiel que l'image de référence de la paire d'images présentée n'ait pas de défauts si l'on veut obtenir des résultats stables.

Les images fixes et les séquences vidéo stockées numériquement sont les plus facilement reproductibles et, partant, les signaux source préférés. Elles peuvent être échangées entre différents laboratoires, cela afin d'obtenir de meilleures comparaisons entre systèmes.

Il faudra souvent tenir compte de l'influence possible, sur la qualité du signal étudié, de tout traitement qui a pu être effectué à un stade antérieur de l'histoire du signal. Par conséquent, quand on procède à des essais sur des sections de la chaîne qui peuvent introduire des distorsions dues au traitement, même si elles sont invisibles, il faut que le signal obtenu soit enregistré de façon transparente et soit

donc disponible pour d'autres essais en aval lorsqu'on veut savoir comment les dégradations dues à une cascade de traitements peuvent s'accumuler le long de la chaîne. Ce genre d'enregistrements sera conservé dans une bibliothèque de matériel d'essai en vue d'une utilisation ultérieure, si besoin est, et on y joindra une description détaillée de l'histoire du signal enregistré. Au besoin, les analyseurs de diapositives 35 mm peuvent être une source d'images fixes. La résolution disponible est satisfaisante pour l'évaluation des systèmes de télévision normale. La colorimétrie et d'autres caractéristiques du film peuvent donner une apparence subjective différente des images de caméra de studio. Si ces paramètres ont une influence sur les résultats, il convient d'utiliser des sources de signaux provenant directement du studio, bien qu'elles soient souvent beaucoup moins pratiques. En règle générale, les analyseurs de diapositives seront adaptés pour chaque image afin d'obtenir la meilleure qualité subjective possible de l'image puisque telle serait la situation dans la pratique.

On évalue souvent les possibilités de post-traitement grâce à la technique d'incrustation d'image. En studio, l'incrustation d'image est très sensible à l'éclairage. Pour les évaluations, il faut donc utiliser de préférence une paire de diapositives spécialement conçues pour l'incrustation d'image qui donnera toujours de très bons résultats. On peut introduire, le cas échéant, un mouvement dans la diapositive de premier plan.

2.3 Choix du matériel d'essai

Plusieurs méthodes ont servi à définir le type de matériel d'essai nécessaire aux évaluations de la télévision. Cependant, dans la pratique, il faut utiliser certains types de matériel d'essai pour traiter des problèmes d'évaluation particuliers. Le Tableau 1-2 indique les problèmes classiques d'évaluation et le matériel d'essai qui sert à les traiter.

TABLEAU 1-2

Choix du matériel d'essai*

Problème d'évaluation	Matériel utilisé
Qualité globale avec matériel moyen	Général, «critique sans excès»
Capacité, applications critiques (par exemple: contribution, post-traitement, etc.)	Gamme étendue, y compris matériel très critique pour l'application à l'essai
Qualité des systèmes «adaptatifs»	Matériel très critique pour le schéma «adaptatif» utilisé
Recenser les points vulnérables et les améliorations possibles	Matériel critique propre à la caractéristique
Identifier les paramètres qui distinguent les systèmes	Large gamme de séquences complexes
Conversion de normes	Critique pour ce qui les distingue (par exemple, fréquence de trame)

* Il va de soi que tout matériel d'essai est du type de ceux qu'on pourrait rencontrer dans des programmes de télévision. Voir les Annexes 3 et 4 pour plus de renseignements sur le choix du matériel d'essai.

Pour certains paramètres, les dégradations observées sur la plupart des images ou des séquences d'images peuvent être plus ou moins identiques. Dans ces conditions, les résultats obtenus à partir d'un très petit nombre d'images ou de séquences d'images (par exemple, 2) peuvent rester significatifs.

Toutefois, l'impact des nouveaux systèmes dépend souvent pour beaucoup de la scène et du contenu de la séquence. En pareil cas, pendant la totalité des heures de programme, il y aura une distribution statistique des probabilités de dégradation et du contenu des images ou des séquences d'images. Étant donné qu'en règle générale on ne connaît pas la forme de cette distribution statistique, le choix du matériel d'évaluation et l'interprétation des résultats doivent être faits avec beaucoup de soin.

En général, il est essentiel d'avoir un matériel de caractère critique car il est possible de tenir compte de ce facteur lors de l'interprétation des résultats mais il n'est pas possible d'extrapoler des résultats à partir d'un matériel non critique. Lorsque la scène ou le contenu de la séquence influence les résultats, il convient de choisir un matériel «critique mais sans excès» pour le système à évaluer. Par «sans excès» on entend que les images pourront raisonnablement faire partie de programmes normaux. En pareil cas, il faut utiliser au moins quatre éléments: par exemple, deux d'entre eux sont véritablement critiques, les deux autres le sont modérément.

2.3.1 Séquences d'essai de l'UIT-R

Un certain nombre d'organisations ont mis au point des images fixes ou des séquences de test. Le Rapport UIT-R BT.2245 sur les matériels d'essai pour la TVHD et la TVUHD, y compris pour la TV-HDR aux fins de l'évaluation de la qualité des images donne des informations détaillées sur les matériels d'essai de TVHD et TVUHD qui peuvent être utilisés pour l'évaluation subjective. Les Annexes 1 et 2 de la Partie 1 de la présente Recommandation donnent des indications supplémentaires pour le choix du matériel d'essai.

2.4 Gamme de conditions et ancrage

Étant donné que la plupart des méthodes d'évaluation sont sensibles aux variations de la gamme et de la distribution des conditions observées, les séances d'évaluation subjective doivent inclure les gammes complètes de variation des facteurs. On peut atteindre toutefois plus ou moins le même objectif avec une gamme plus restreinte en présentant également certaines conditions qui se situeront aux extrémités des échelles. Elles peuvent être représentées comme exemples et identifiées comme étant les plus extrêmes (ancrage direct) ou réparties tout au long de la séance et non identifiées comme étant les plus extrêmes (ancrage indirect).

2.5 Observateurs

Les observateurs peuvent être spécialistes ou non, suivant les objectifs de l'évaluation. Un observateur spécialiste est un observateur qui a des connaissances spécialisées sur les artéfacts susceptibles d'être introduits dans les images par le système à évaluer. Un observateur non spécialiste («novice») est un observateur qui n'a pas de connaissances spécialisées sur les artéfacts susceptibles d'être introduits dans les images par le système à évaluer. Dans tous les cas, les observateurs ne doivent pas participer ou avoir participé directement – à savoir suffisamment pour acquérir des connaissances spécifiques et détaillées – à la mise au point du système à évaluer.

2.5.1 Nombre d'observateurs

Sauf indication contraire de la méthodologie choisie, il faut au moins 15 observateurs. Le nombre d'observateurs dépend de la sensibilité et de la fiabilité de la procédure d'essai retenue ainsi que de l'ampleur escomptée de l'effet évalué. Pour les études de portée limitée, par exemple des études préliminaires, on pourra recourir à moins de 15 observateurs. Dans ce cas, l'étude devra être identifiée comme étant «informelle». Le niveau de compétence des observateurs en matière d'évaluation de la qualité d'images de télévision devra être mentionné.

2.5.2 Sélection des observateurs

En général, avant chaque séance, les observateurs seront sélectionnés à l'aide de mires de Snellen ou de Landolt pour leur acuité visuelle normale ou rendue normale par correction et leur vision normale des couleurs, cela à l'aide de mires choisies à cet effet (d'Ishihara, par exemple).

On trouvera dans les §§ A1-2.3 et A1-2.4 des informations détaillées sur différents scénarios de sélection des observateurs qui peuvent être appliqués à différentes méthodologies d'essai. Si des essais en laboratoire ou des essais moins formels sont menés dans le cadre d'un programme d'essai multi-

sites ou d'une organisation, il est important que toutes les informations relatives à la méthode et aux critères de sélection des observateurs soient communiquées et incluses dans les résultats publiés.

En général, il faut donner le plus de détails possible sur les caractéristiques des groupes d'évaluation qui ont été retenus, telles que des précisions sur l'activité professionnelle (fonctionnaire d'un organisme de radiodiffusion, étudiant d'une université, personnel de bureau, par exemple), le sexe et l'âge.

NOTE – Une étude destinée à vérifier la cohérence des résultats obtenus par des laboratoires d'essai différents a permis de constater l'existence possible de différences systématiques entre ces résultats, différences qui seront particulièrement importantes s'il est proposé de regrouper les résultats de plusieurs laboratoires différents afin d'améliorer la sensibilité et la fiabilité d'une expérience.

Ces différences peuvent s'expliquer par le fait que des groupes d'observateurs différents peuvent avoir des niveaux d'aptitude différents. Il faut entreprendre des recherches plus poussées pour voir si cette hypothèse se confirme et, dans l'affirmative, pour mesurer les variations dues à ce facteur.

2.5.3 Instructions pour les évaluations

La méthode d'évaluation, les types de dégradation ou de facteur de qualité auxquels il faut s'attendre, l'échelle d'évaluation, la séquence elle-même et le séquencement seront présentés avec soin aux observateurs. Les séquences d'entraînement qui montrent la gamme et le type de dégradation à évaluer doivent présenter d'autres images que celles qui sont utilisées dans les essais mais avoir une sensibilité comparable. Dans les évaluations de la qualité, cette dernière peut être définie par des propriétés perceptuelles spécifiques.

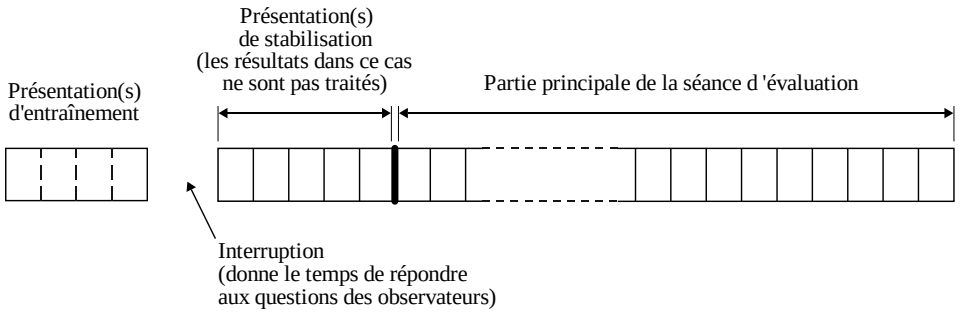
2.6 Séance d'évaluation

Une séance ne devrait pas dépasser une demi-heure. Au début de la première séance, on procédera à environ cinq «présentations fictives» pour stabiliser les jugements des observateurs. On ne tiendra pas compte de leurs résultats dans le dépouillement des essais. Si plusieurs séances sont nécessaires, seules trois présentations fictives environ sont nécessaires au début de la séance suivante.

Il convient de choisir un ordre aléatoire pour la présentation des images (par exemple, déduit de carrés gréco-latins); quoi qu'il en soit, les conditions d'essai doivent être présentées dans un ordre permettant d'équilibrer, séance après séance, tous les effets que les phénomènes de fatigue et d'adaptation peuvent avoir sur les notations. Certaines présentations peuvent être répétées d'une séance à l'autre pour vérifier la cohérence.

FIGURE 1-2

Structure de la présentation d'une séance d'évaluation



BT.0500-01-2

2.7 Présentation des résultats

Étant donné que les résultats varient en fonction de la gamme, il n'est pas judicieux d'interpréter en termes absolus les évaluations obtenues à partir de la plupart des méthodes (par exemple, la qualité d'une image ou d'une séquence d'images).

Pour chaque paramètre d'essai, il faut donner la moyenne et l'intervalle de confiance à 95% de la distribution statistique des notes d'évaluation. Si l'évaluation portait sur la variation de la dégradation en fonction de la variation de la valeur d'un paramètre, il conviendra d'utiliser des courbes de régression. Une courbe de régression appropriée en coordonnées logarithmiques permettra de représenter les résultats sous forme d'une droite. C'est là le mode de présentation préféré. On trouvera dans l'Annexe 1 de la Partie 1 de la présente Recommandation des informations supplémentaires sur le traitement des données.

Les résultats doivent être donnés avec les informations suivantes:

– détails de la configuration de l'expérience;
– détails du matériel d'évaluation;
– type de source d'image et d'écran de visualisation (voir la Note 1);
– nombre et type d'observateurs (voir la Note 2);
– systèmes de référence utilisés;
– moyenne générale de l'expérience;
– résultats moyens originaux et corrigés et intervalle de confiance à 95% si un ou plusieurs observateurs ont été éliminés selon la procédure ci-dessous.

NOTE 1 – Étant donné que certains éléments donnent à penser que la taille de l'écran peut avoir une incidence sur les résultats des évaluations subjectives, il est demandé aux expérimentateurs d'indiquer explicitement la taille de l'écran ainsi que la marque et le numéro de modèle des systèmes de visualisation utilisés dans les différentes expériences.

NOTE 2 – Il apparaît que les différences d'aptitude entre groupes d'observateurs (même entre groupes «non spécialistes») peuvent avoir une incidence sur les résultats des évaluations subjectives. Pour faciliter l'étude de ce phénomène, il est demandé aux expérimentateurs de donner le plus de détails possible sur les caractéristiques des groupes, qu'ils ont retenus en particulier sur l'âge, le sexe, le niveau d'étude ou l'activité professionnelle des différents membres de chaque groupe.

3 Choix des méthodes d'essai

Pour les évaluations de la télévision, on a recouru à des méthodes d'essai fondamentales très diverses. Cependant, dans la pratique, chaque problème d'évaluation particulier suppose des méthodes particulières. La Partie 3 de la présente Recommandation donne des orientations concernant l'évaluation subjective de la qualité des images pour différents formats et applications.

Annexe 1
de la Partie 1

Analyse et présentation des résultats

A1-1 Introduction

Au cours d'expériences subjectives effectuées en vue d'estimer la qualité d'un système de télévision, des données sont rassemblées en grand nombre. Elles se présentent sous forme de feuilles de notes remplies par des observateurs ou leur équivalent électronique et il faut, selon des méthodes statistiques, les concentrer sous une forme graphique et/ou numérique/formules/algorithme qui résume la qualité du système étudié.

L'analyse suivante s'applique aux résultats des méthodes à double stimulus, la méthode DSIS et la méthode DSCQS qui servent toutes deux à évaluer la qualité des images de télévision et qui sont décrites dans les Annexes 1, 2 et 3 de la Partie 2 de la présente Recommandation, ainsi qu'à d'autres méthodes avec échelles numériques. Pour la première et la seconde de ces méthodes, la dégradation est notée sur une échelle à 5 notes ou une échelle multinote. Pour les dernières, on utilise des échelles de notation continue et les résultats (différence entre les notes de l'image de référence et de l'image soumise aux essais) sont normalisés à une valeur entière comprise entre 0 et 100.

A1-2 Méthodes communes d'analyse

Les tests effectués conformément aux principes des méthodes décrites au § 2 de la Partie 1 aboutiront à des distributions d'entiers compris par exemple entre 1 et 5 ou 0 et 100, qui varieront en raison des différences d'évaluation entre les observateurs et de l'influence de divers paramètres liés à l'expérience, par exemple l'utilisation de plusieurs images ou séquences.

Un test se composera d'un certain nombre, *L*, de présentations, chacune étant constituée d'un certain nombre, *J*, de conditions de test appliquées à l'une des *K* séquences de test/images test. Dans certains cas, chaque combinaison de séquences de test/image test et de condition de test peut être répétée *R* fois.

A1-2.1 Calcul des notes moyennes

La première étape de l'analyse des résultats est le calcul de la note moyenne, $\bar{u}_{jkr}$, pour chacune des présentations:

$$\bar{u}_{jkr} = \frac{1}{N} \sum_{i=1}^{N} u_{ijkr} \tag{1}$$

où:

u_{ijkr}: note de l'observateur *i* pour la condition de test *j*, la séquence/image *k* et la répétition *r*

N: nombre d'observateurs.

On pourrait de même calculer les notes moyennes d'ensemble, $\bar{u}_j$ et $\bar{u}_k$, pour chaque condition et séquence/image de test.

A1-2.2 Calcul de l'intervalle de confiance

A1-2.2.1 Traitement de données brutes (n'ayant fait l'objet d'aucune compensation ni d'aucune approximation)

Lors de la présentation des résultats d'un test, on associera à toutes les notes moyennes un intervalle de confiance calculé à partir de l'écart type et de la taille de chaque échantillon.

Il est proposé d'utiliser l'intervalle de confiance à 95% donné par:

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

où:

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

L'écart type de chaque présentation, S_{jkr}, est donné par:

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

Avec une probabilité de 95%, la valeur absolue de la différence entre la note moyenne expérimentale et la note moyenne «réelle» (pour un très grand nombre d'observateurs) est inférieure à l'intervalle de confiance de 95%, sous réserve que la distribution des différentes notes remplisse certaines conditions.

On pourrait de même calculer un écart type, S_j, pour chaque condition de test. On notera toutefois que, lorsque le nombre de séquences de test est faible, cet écart type sera davantage influencé par les différences entre les séquences/images de test utilisées que par les différences d'évaluation entre les observateurs participant à l'évaluation.

A1-2.2.2 Traitement de données ayant fait l'objet d'une compensation ou d'une approximation

Pour les données pour lesquelles les effets des dégradations/améliorations résiduelles ou les effets de fin d'échelle sur les échelles d'évaluation ont été compensés ou pour les données présentées sous la forme de réaction aux dégradations ou de loi d'addition des dégradations après approximation, données qui influent sur les notes moyennes de qualité obtenues expérimentalement, il faut calculer l'intervalle de confiance au moyen de transformations des variables statistiques en tenant compte de la dispersion des valeurs de ces variables.

Si les résultats de l'évaluation de la qualité sont présentés sous la forme d'une réaction aux dégradations (c'est-à-dire d'une courbe expérimentale), les limites supérieure et inférieure de l'intervalle de confiance seront fonction de chaque valeur expérimentale. Pour déterminer ces limites, il faut calculer l'écart type et évaluer par approximation, pour chaque valeur expérimentale de la réaction initiale aux dégradations, son influence.

A1-2.3 Sélection a posteriori des observateurs

A1-2.3.1 Sélection a posteriori fondée sur le kurtosis pour les méthodes DSIS, DSCQS et les autres méthodes, à l'exception de la méthode SSCQE

Il faut d'abord vérifier, au moyen du test β_2, si la distribution des notes pour chaque présentation est normale ou non (en calculant le coefficient d'aplatissement (kurtosis) de la fonction, c'est-à-dire le rapport du moment d'ordre quatre sur le carré du moment d'ordre deux. Si β_2 est compris entre 2 et 4, on peut considérer que la distribution est normale. Pour chaque présentation, les notes u_{ijkr} de chaque

observateur doivent être comparées à la valeur moyenne associée $\bar{u}_{jkr}$ plus l'écart type associé S_{jkr} multiplié par 2 (normale) ou par $\sqrt{20}$ (non normale), P_{jkr}, et à la valeur moyenne associée, moins ce même écart type multiplié par 2 ou $\sqrt{20}$, Q_{jkr}. Chaque fois qu'une note donnée par un observateur est supérieure à P_{jkr}, un compteur associé à chaque observateur P_i est incrémenté. De même, chaque fois qu'une note donnée par un observateur est inférieure à Q_{jkr}, un compteur associé à chaque observateur Q_i est incrémenté. On doit enfin calculer les deux rapports suivants: $P_i + Q_i$ sur le nombre total de notes données par chaque observateur au cours de toute la séance et $P_i - Q_i$ sur $P_i + Q_i$ en valeur absolue. Si le premier est supérieur à 5% et le second inférieur à 30%, il faut éliminer l'observateur *i* (voir la Note).

NOTE – Il ne faut pas appliquer cette procédure plus d'une fois aux résultats d'une expérience donnée. En outre, on la réservera aux cas où il y a relativement peu d'observateurs (moins de 20 par exemple) et aucun spécialiste.

Il est recommandé d'utiliser cette procédure pour la méthode de l'UER (DSIS); cette procédure a également été appliquée avec succès à la méthode DSCQS et à d'autres méthodes.

Le processus ci-dessus peut s'exprimer mathématiquement comme suit:

Pour chaque présentation de test, calculer la moyenne $\bar{u}_{jkr}$, l'écart type S_{jkr}, et le coefficient de kurtosis β_{2jkr}, où β_{2jkr} est donné par:

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

Pour chaque observateur *i* trouver P_i et Q_i, c'est-à-dire:

pour *j*, *k*, *r* = 1, 1, 1 à *J*, *K*, *R*

si $2 \le \beta_{2jkr} \le 4$, alors:

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

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

sinon:

si $u_{ijkr} \ge \bar{u}_{jkr} + \sqrt{20}\,S_{jkr}$ alors $P_i = P_i + 1$

si $u_{ijkr} \le \bar{u}_{jkr} - \sqrt{20}\,S_{jkr}$ alors $Q_i = Q_i + 1$

Si $\dfrac{P_i + Q_i}{J \cdot K \cdot R} > 0{,}05$ et $\left|\dfrac{P_i - Q_i}{P_i + Q_i}\right| < 0{,}3$ alors rejeter l'observateur *i*

avec:

N: nombre d'observateurs

J: nombre de conditions de test y compris la référence

K: nombre d'images ou de séquences de test

R: nombre de répétitions

L: nombre de présentations de test (dans la plupart des cas, le nombre de présentations sera égal à $J \cdot K \cdot R$, mais on notera que certaines évaluations peuvent être faites avec un nombre inégal de séquences pour chaque condition de test).

A1-2.3.2 Sélection a posteriori fondée sur le kurtosis pour la méthode SSCQE

Pour la sélection spécifique des observateurs en cas d'utilisation de la procédure de test SSCQE, le domaine d'application n'est plus une des configurations de test (combinaison d'une condition de test et d'une séquence de test) mais une fenêtre temporelle (par exemple, un segment de test de 10 s) d'une configuration de test. On applique un filtrage en deux temps: dans une première étape, le filtrage a pour objet de déceler et d'exclure les observateurs dont les notes sont très décalées par rapport au comportement moyen; dans une seconde étape, l'opération vise à déceler et à filtrer les observateurs «irréguliers» sans tenir compte de décalages systématiques.

Étape 1: Détection des inversions de note locales

Ici également, il faut d'abord vérifier, au moyen du test β_2, si la distribution des notes pour chaque fenêtre temporelle de chaque configuration de test est «normale» ou non. Si β_2 est compris entre 2 et 4, on peut considérer que la distribution est «normale». Le processus s'applique alors à chaque fenêtre temporelle de chaque configuration de test et son traitement mathématique est indiqué ci-après.

Pour chaque fenêtre temporelle de chaque configuration de test, et en utilisant les notes u_{ijkr} de chaque observateur, calculer la moyenne $\bar{u}_{jklr}$, l'écart type S_{jklr}, et le coefficient de β_{2jklr}, où β_{2jklr} est donné par:

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

Pour chaque observateur, i, trouver P_i et Q_i, c'est-à-dire:

pour j, k, l, r = 1,1,1,1 à J, K, L, R

si $2 \le \beta_{2jklr} \le 4$, alors:

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

si $u_{njklr} \le \bar{u}_{jklr} + 2\,S_{jklr}$ alors $Q_i = Q_i + 1$

sinon:

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

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

Si $\frac{P_i}{J \cdot K \cdot L \cdot R} > X\%$ ou $\frac{Q_i}{J \cdot K \cdot L \cdot R} > X\%$ alors rejeter l'observateur i

avec:

N: nombre d'observateurs

J: nombre de fenêtres temporelles dans une combinaison de test (condition de test + séquence de test)

K: nombre de conditions de test

L: nombre de séquences

R: nombre de répétitions.

Ce processus permet de rejeter les observateurs dont les notes étaient très éloignées des notes moyennes. La Figure 1-3 en montre deux exemples (les deux courbes extrêmes mettent en évidence de grands décalages). Néanmoins, ce critère de rejet ne permet pas de détecter les inversions possibles, qui sont une autre source importante d'erreurs systématiques sur les résultats. Pour cette raison, une seconde étape est proposée.

Étape 2: Détection des inversions de note locales

Dans l'Étape 2, la détection se fonde également sur les formules de sélection données dans la présente Annexe avec une légère modification du domaine d'application. Ici encore, l'ensemble des données d'entrée est constitué par les notes afférentes à toutes les fenêtres temporelles (par exemple, 10 s) de toutes les configurations de test. Mais les notes subissent cette fois un traitement préliminaire, à savoir un centrage autour de la moyenne générale, le but étant de minimiser l'effet de décalage qui a déjà été traité dans la première étape. On procède ensuite à l'application du traitement habituel.

Il faut d'abord vérifier, au moyen du test β_2, si cette distribution des notes pour chaque fenêtre temporelle de chaque configuration de test est «normale» ou non. Si β_2 est compris entre 2 et 4, on peut considérer que la distribution est normale. Le processus s'applique alors à chaque fenêtre temporelle de chaque configuration de test et son traitement mathématique est indiqué ci-après.

La première opération du processus est le calcul des notes centrées pour chaque fenêtre temporelle et chaque observateur. Pour chacune des configurations de test, la note moyenne, $\bar{u}_{klr}$ est définie par:

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

De la même façon, pour chaque configuration de test et chaque observateur, la note moyenne est définie par:

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

u_{njklr} correspond à la note de l'observateur i pour la fenêtre temporelle j, la condition de test k, la séquence l et la répétition r.

Pour chaque observateur, les notes centrées u^*_{njklr} se calculent comme suit:

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

Pour chaque fenêtre temporelle de chaque configuration de test, on calcule la moyenne $\bar{u}^*_{jklr}$, l'écart type S^*_{jklr}, et le coefficient β_{2jklr}, ce dernier coefficient étant donné par:

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

Pour chaque observateur, i, trouver P^*_i et Q^*_i, c'est-à-dire:

Pour j, k, l, r = 1, 1, 1, 1 à J, K, L, R

si $2 \le \beta_2{}^*_{jklr} \le 4$, alors:

si $u^*_{njklr} \ge \bar{u}^*_{jklr} + 2\, S^*_{jklr}$ alors $P^*_i = P^*_i + 1$

si $u^*_{njklr} \le \bar{u}^*_{jklr} - 2\, S^*_{jklr}$ alors $Q^*_i = Q^*_i + 1$

sinon:

si $u^*_{njklr} \geq \bar{u}^*_{jklr} + \sqrt{20}\ S^*_{jklr}$ alors $P^*_i = P^*_i + 1$

si $u^*_{njklr} \leq \bar{u}^*_{jklr} - \sqrt{20}\ S^*_{jklr}$ alors $Q^*_i = Q^*_i + 1$

Si $\dfrac{P^*_i + Q^*_i}{J \cdot K \cdot L \cdot R} > Y$ et $\left|\dfrac{P^*_i - Q^*_i}{P^*_i + Q^*_i}\right| < Z$ alors rejeter l'observateur i

avec:

N: nombre d'observateurs

J: nombre de fenêtres temporelles dans une combinaison de test (condition de test + séquence de test)

K: nombre de conditions de test

L: nombre de séquences

R: nombre de répétitions.

L'expérience montre que les valeurs à proposer pour les paramètres (X, Y, Z) adaptés à cette méthode sont 0,2, 0,1 et 0,3.

A1-2.3.3 Sélection a posteriori fondée sur la corrélation

Chaque observateur doit disposer d'une méthode stable et cohérente pour évaluer correctement la dégradation de qualité pour chaque scène et algorithme. Le critère de rejet permet de confirmer le niveau de cohérence des notes d'un observateur par rapport à la note moyenne pour tous les observateurs ayant participé à une séance d'essai donnée. Le critère de décision est fondé sur une corrélation des notes individuelles avec les notes moyennes correspondantes pour tous les observateurs ayant participé à l'essai. La procédure est plus simple à mettre en œuvre que la méthode correspondante décrite dans les paragraphes précédents.

A1-2.3.3.1 Corrélation de Pearson

La relation entre l'échelle de qualité et la plage de notes données par les observateurs est supposée être linéaire pour pouvoir appliquer la corrélation de Pearson.

Le principal objectif est de vérifier par une méthode simple si les notes données par un observateur sont cohérentes avec les notes moyennes pour tous les observateurs et pour l'ensemble de la séance d'essai. La référence cachée est considérée comme une ancre de qualité élevée. Si les ancres faible et élevée sont incluses, elles augmentent la note de corrélation; inversement, les décalages de corrélation entre les observateurs sont réduits.

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

où:

x_i: note moyenne pour tous les observateurs pour le triplet (algorithme, débit, scène)

y_i: note donnée par un seul observateur pour le même triplet

n: (nombre d'algorithmes) × (nombre de scènes)

i: {nombre de codec, nombre de débit, nombre de scène}.

A1-2.3.3.2 Corrélation des rangs de Spearman

La corrélation des rangs de Spearman peut être appliquée même si la relation entre l'échelle de qualité et la plage de notes données par les observateurs n'est pas supposée être linéaire[4]:

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

où:

x_i: note moyenne pour tous les observateurs pour le triplet (algorithme, débit, scène)

y_i: note donnée par un seul observateur pour le même triplet

n: (nombre d'algorithmes) × (nombre de scènes)

$R(x_i$ ou $y_i)$: rang

i: {nombre de codec, nombre de débit, nombre de scène}.

A1-2.3.3.3 Critères de rejet final pour éliminer un observateur ayant participé à un essai donné

La corrélation des rangs de Spearman et la corrélation de Pearson sont utilisées pour éliminer un ou plusieurs observateurs conformément aux conditions suivantes:

SI [moyenne(r) – sdt(r)] > Seuil de corrélation maximal (MCT).
Seuil de rejet = Seuil de corrélation maximal (MCT).
SINON Seuil de rejet = [moyenne(r) – écart type(r)].

SI [r (Observateur$_i$)] > Seuil de rejet.
ALORS l'observateur «i» ayant participé à l'essai n'est pas éliminé.
SINON l'observateur «i» ayant participé à l'essai est éliminé.

où:

r = min (corrélation de Pearson, corrélation des rangs de Spearman)

moyenne(r): moyenne des corrélations pour tous les observateurs ayant participé à un essai donné

sdt(r): écart type des corrélations pour tous les observateurs ayant participé à un essai donné

Seuil de corrélation maximal (MCT) = 0,85.

La valeur de MCT de 0,85 est valable pour les méthodes SAMVIQ et DSCQS; une valeur de MCT de 0,7 doit être utilisée pour les méthodes à un seul stimulus et DSIS.

---

[4] Les résultats de corrélation de Pearson sont généralement très proches des résultats de corrélation de Spearman.

A1-2.4 Calcul des notes moyennes et des intervalles de confiance dans des conditions d'essai difficiles

Un essai subjectif doit très souvent être mené dans des conditions difficiles. Par exemple, dans le cadre d'un essai participatif, les participants sont exposés à un environnement moins contrôlé que celui d'un laboratoire. Dans le cadre d'un essai à grande échelle mené par plusieurs laboratoires, la variabilité inter-laboratoire pourrait donner lieu à une large variance des notes recueillies. Les méthodes présentées dans les §§ A1-2.1 à A1-2.3 ne sont généralement pas adaptées à ces situations. Il est ici question de présenter une technique d'analyse de données avancée dont il a été démontré qu'elle améliorait la qualité des données des notes moyennes et des intervalles de confiance obtenues. On trouvera également une mise en œuvre Python de référence dans la Pièce jointe 1 à la présente Annexe.

L'idée qui sous-tend cette technique est la suivante. Il est utile de modéliser explicitement le comportement de chaque participant; les biais et la cohérence d'un participant, en particulier, sont deux facteurs humains importants qui influent sur les notes données par l'intéressé. Par l'intermédiaire d'une procédure itérative, cette technique tente d'estimer conjointement la qualité réelle de chaque présentation et les biais et la cohérence de chaque participant. La qualité réelle estimée de chaque présentation peut être interprétée comme étant une «note d'opinion moyenne dénuée de biais avec pondération de la cohérence». En comparaison avec la sélection a posteriori des participants décrite au § A1-2.3.1, dans laquelle toutes les notes des intéressés sont conservées ou rejetées («rejet ferme»), cette technique peut être décrite comme une technique de «rejet modéré». En d'autres termes, les notes incohérentes données par un participant qui fait figure d'exception auraient peu de poids et influeraient donc peu sur la note moyenne d'opinion globale. Cette technique a pour conséquence de devoir estimer les biais et la cohérence de chaque participant à l'essai. Il s'agit d'informations utiles permettant de déterminer si la participation d'une personne à des essais subjectifs est pertinente; par conséquent, elles peuvent être utilisées pour sélectionner des participants en vue de futurs essais. Par exemple, si l'on constate que l'un d'eux note de manière très incohérente, on pourra éventuellement l'empêcher de participer à des sessions futures.

Dans cette technique, on estime d'abord les notes moyennes pour chaque présentation parmi l'ensemble des participants et des répétitions:

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

où u_{ijkr} est la note de l'observateur *i* pour la condition *j*, la séquence/l'image *k*, la répétition *r*, *N* étant le nombre d'observateurs et *R* le nombre de répétitions.

Lors de la deuxième étape, les biais de chaque observateur b_i sont estimés comme suit:

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

J et *K* étant respectivement le nombre de conditions et le nombre de séquences. Les étapes suivantes sont alors menées dans une boucle itérative.

L'estimation actuelle de la note moyenne pour chaque présentation est écrite $\bar{u}_{jk}^{c}$, c'est-à-dire que:

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

On calcule alors les résidus dans chaque note observée ne pouvant être expliquée par la note moyenne et les biais de l'observateur:

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

On utilise alors ces résidus pour calculer l'incohérence de chaque observateur σ_i comme suit:

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

où:

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

Les nouvelles estimations des notes moyennes peuvent alors être obtenues comme suit:

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

On actualise alors les biais en fonction de l'équation (12).

La boucle est terminée si:

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

Une fois la boucle terminée, l'écart type de la note pour chaque présentation est obtenu comme suit:

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

où:

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

et

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

L'intervalle de confiance final est alors calculé en fonction des équations (2) et (3).

A1-3 Méthode permettant de trouver une correspondance entre la note moyenne et la mesure objective d'une distorsion de l'image

Si les essais subjectifs ont été effectués pour étudier la relation entre la mesure objective d'une distorsion et les notes moyennes $\bar{u}$ ($\bar{u}$ calculé comme indiqué au § A1-2.1), la méthode suivante peut être utile; elle consiste à rechercher une correspondance simple entre $\bar{u}$ et le paramètre dégradation.

A1-3.1 Approximation par une fonction logistique symétrique

L'approximation de cette relation expérimentale par une fonction logistique se révèle particulièrement intéressante.

Le traitement des données $\bar{u}$ peut se faire comme suit:

L'échelle des valeurs de $\bar{u}$ est normalisée par référence à une variable continue *p* telle que,

$$p = (\bar{u} - u_{min})/(u_{max} - u_{min}) \tag{24}$$

avec:

u_{min}: note minimum disponible sur l'échelle des *u* pour la moins bonne qualité

u_{max}: note maximum disponible sur l'échelle des *u* pour la meilleure qualité.

La représentation graphique de la relation entre *p* et *D* montre que la courbe peut avoir une allure de sigmoïde à symétrie centrale si les limites naturelles des valeurs de *D* sont très éloignées du domaine dans lequel *u* varie rapidement.

La fonction $p = f(D)$ se prête alors à une approximation par une fonction logistique judicieusement choisie, répondant à la relation générale:

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

où D_M et G sont constants et où G peut être positif ou négatif.

La valeur p répondant à la fonction logistique d'approximation optimale est utilisée pour fournir une valeur numérique déduite, I, répondant à la relation:

$$I = (1/p - 1) \quad (26)$$

Les valeurs de D_M et de G peuvent s'obtenir à partir des données expérimentales après la transformation suivante:

$$I = \exp(D - D_M) \cdot G \quad (27)$$

En portant I sur une échelle logarithmique, on obtient alors une relation linéaire:

$$\log_e I = (D - D_M) \cdot G \quad (28)$$

L'interpolation par une droite est alors aisée et, dans certains cas, d'une précision suffisante pour que cette droite puisse être considérée comme représentant la dégradation due à l'effet mesuré par D.

La pente de la caractéristique s'exprime alors par:

$$S = \frac{D_M - D}{\log_e I} = \frac{1}{G} \quad (29)$$

ce qui fournit la valeur optimale de G. D_M est la valeur de D pour $I = 1$.

La droite constitue la caractéristique de dégradation, associée à l'effet dégradant considéré. On notera que la droite peut être définie par les valeurs caractéristiques D_M et G de la fonction logistique.

A1-3.2 Approximation par une fonction non symétrique

A1-3.2.1 Description de la fonction

L'utilisation d'une fonction logistique symétrique pour approximer la relation entre les notes expérimentales et la mesure objective d'une distorsion de l'image donne les meilleurs résultats lorsque le paramètre de distorsion D peut être mesuré dans une unité connexe, par exemple le rapport S/N (dB). Si ce paramètre est mesuré dans une unité physique d, par exemple un délai (ms), la relation (27) doit être remplacée par:

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

ce qui donne pour l'équation (25):

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

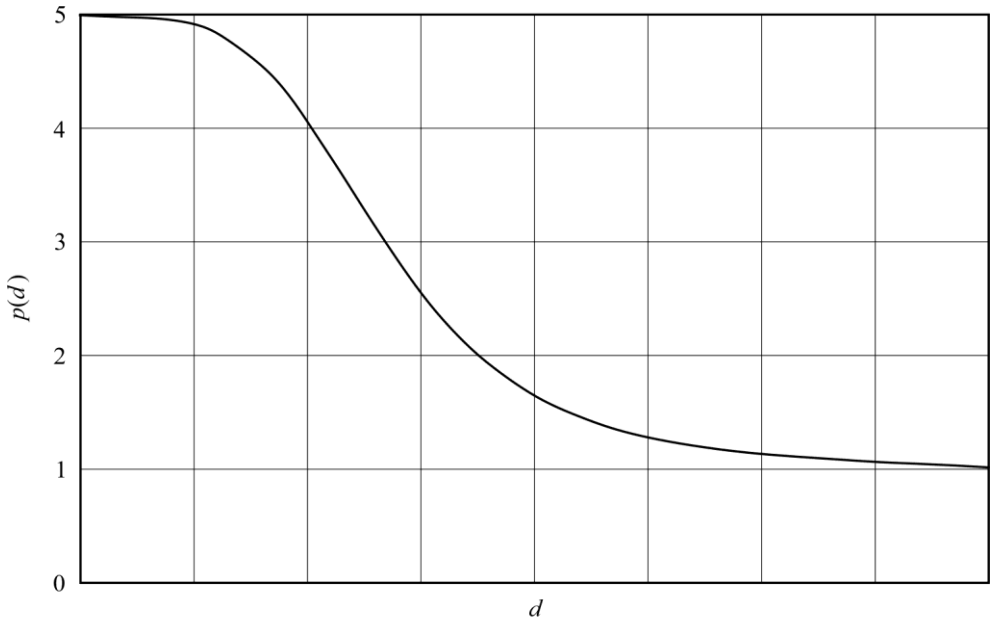
Cette fonction constitue une approximation non symétrique de la fonction logistique.

A1-3.2.2 Estimation des paramètres de l'approximation

L'estimation des paramètres optimaux de la fonction qui donne le moins d'erreurs résiduelles entre les données effectives et la fonction peut être obtenue avec n'importe quel algorithme d'estimation récursive. La Figure 1-3 montre un exemple de l'utilisation de la fonction non symétrique pour représenter des données subjectives réelles. Cette représentation permet l'estimation des mesures objectives spécifiques correspondant à une valeur subjective intéressante: par exemple 4,5 sur l'échelle à cinq notes.

FIGURE 1-3

Approximation non symétrique



BT.0500-01-3

A1-3.3 Correction de la dégradation/amélioration résiduelle et de l'effet de fin d'échelle

Dans la pratique, l'emploi d'une fonction logistique ne permet pas toujours d'éviter des différences entre les données expérimentales et l'approximation. Ces différences peuvent être dues aux effets de fin d'échelle ou à la présence simultanée de plusieurs dégradations dans le test, ce qui peut influencer le modèle statistique et déformer la fonction logistique théorique.

On a décelé une sorte d'effet de fin d'échelle, à savoir que les observateurs évitent d'utiliser les valeurs extrêmes de l'échelle d'évaluation, en particulier pour les notes de qualité élevées. Il peut y avoir à cela plusieurs raisons, comme une répugnance psychologique à porter des jugements extrêmes. Par ailleurs, l'utilisation de la moyenne arithmétique des jugements selon l'équation (1) au voisinage des extrémités de l'échelle peut conduire à des résultats faussés, en raison de la distribution non gaussienne des notes dans ces régions.

On indique fréquemment dans les tests une «dégradation résiduelle» (même dans les images de référence, la note moyenne atteint seulement une valeur $\bar{u}_0 < u_{max}$).

Il existe un certain nombre de méthodes utiles pour corriger les données brutes des évaluations afin d'aboutir à des conclusions valables (voir le Tableau 1-3).

La correction des effets de fin d'échelle, si ceux-ci sont présents dans les données expérimentales, est une partie très importante du traitement des données. Il faut par conséquent choisir la procédure avec le plus grand soin. Il convient de signaler que ces procédures de correction reposent sur des hypothèses spéciales. Il est donc conseillé d'agir avec prudence lorsqu'on emploie lesdites procédures dont l'utilisation doit être signalée dans la présentation des résultats.

TABLEAU 1-3

Comparaison des méthodes de correction des effets de fin d'échelle

Méthodes par compensation des effets de fin d'échelle	Caractéristiques		
	Compensation de la dégradation résiduelle	Compensation du renforcement résiduel	Décalage du centre de l'échelle
Pas de compensation	Non	Non	Non
Transformation linéaire de l'échelle	Oui	Peut-être une erreur significative	Non
Transformation non linéaire de l'échelle[1)]	Oui	Oui	Non
Méthode fondée sur l'addition des unités imp	Oui	Non	Oui
Méthode multiplicative	Oui	Non	Oui

1) Dans la transformation non linéaire de l'échelle, il faut calculer les notes corrigées:

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

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

avec:

u_{corr} : note corrigée

$\bar{u}$: note expérimentale non corrigée

u_{min}, u_{max} : limites de l'échelle d'évaluation

u_{mid} : point milieu de l'échelle d'évaluation

$u_{0\ min}$, $u_{0\ max}$: limites inférieure et supérieure de la tendance des notes expérimentales.

A1-3.4 Incorporation de l'aspect fiabilité dans les graphiques

À partir des notes moyennes de chaque dégradation testée et de l'intervalle de confiance à 95% associé, on construit trois séries de notes:

– série de notes minimales (moyennes – intervalles de confiance);

– série de notes moyennes;

– série de notes maximales (moyennes + intervalles de confiance).

On procède alors à une estimation des paramètres pour les trois séries indépendamment. Ce qui permet de tracer les trois fonctions obtenues sur le même graphique. Les deux fonctions issues des séries maximales et minimales en pointillés, l'estimation moyenne en trait plein. On pointe aussi sur ce graphique les valeurs expérimentales (voir la Fig. 1-4). On obtient ainsi une estimation de la zone de confiance continue à 95%.

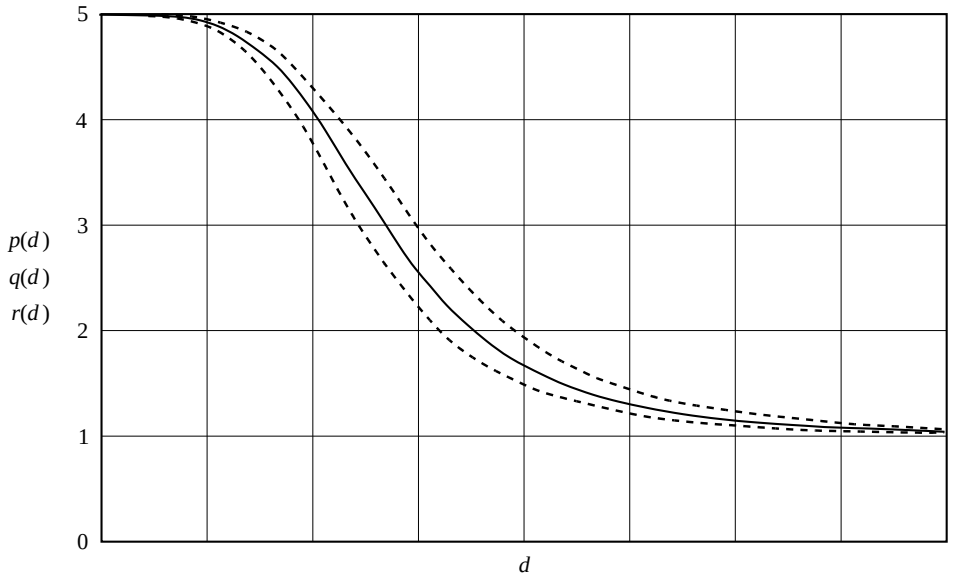
Pour la note 4,5 (seuil de visibilité associé à la méthode), on a donc directement par lecture graphique un intervalle de confiance à 95% estimé pouvant servir à la détermination d'une fourchette de tolérance.

L'espace entre les courbes maximales et minimales n'est pas un intervalle à 95%, mais une estimation moyenne de celui-ci.

Les valeurs expérimentales (au moins 95%) devraient être incluses dans la zone de confiance, sinon on peut penser qu'il y a eu un problème dans le déroulement du test ou que le modèle de fonction choisi n'est pas optimum.

FIGURE 1-4

Cas d'une caractéristique de dégradation non symétrique



p(*d*): série des qualités moyennes
q(*d*): série des qualités minimales
r(*d*): série des qualités maximales
d: mesure objective de la dégradation

BT.0500-01-4

A1-4 Conclusions

On a décrit un processus d'évaluation des intervalles de confiance, c'est-à-dire la précision d'un ensemble d'essais d'évaluations subjectives.

Le processus aboutit aussi à l'estimation de quantités moyennes générales qui ne sont pas restreintes à l'expérience en question mais s'étendent aussi aux autres expériences effectuées avec la même méthodologie.

Ces quantités peuvent donc servir à tracer des diagrammes de comportement de l'intervalle de confiance, ce qui sert aux estimations subjectives ainsi qu'à l'organisation de futures expériences.

Pièce jointe 1
à l'Annexe 1

Mise en œuvre de référence de la méthode décrite au § A1-2.4

La présente pièce jointe comprend une mise en œuvre Python de référence de la technique d'analyse des données présentée au § A1-2.4. Les exemples de code et les données employées sont mis à la disposition du public par l'intermédiaire du paquetage Python SUREAL disponible à l'adresse: https://github.com/Netflix/sureal/tree/master/itur_bt500_demo.

Les données d'entrée sont préparées de la façon suivante: les votes bruts sont agencés sous la forme d'une matrice 2D, dans laquelle ils sont séparés par des virgules. Chaque ligne correspond à une présentation (image source visionnée dans une condition de test); chaque colonne correspond à un participant.

Il n'est pas nécessaire que chaque participant émette un vote pour chaque présentation. Si le participant *i* n'émet pas de vote pour une présentation *jk*, la valeur «nan» (pas un nombre, *not a number*) est insérée à l'emplacement (*jk*,*i*). Les données d'entrée sont placées dans un fichier .csv. On trouvera ci-dessous un exemple de petit fichier .csv contenant les votes de 20 participants sur un ensemble de 30 présentations des images sources, répétées une fois.

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

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

Le code Python utilisé pour mettre en œuvre la méthode se trouve dans le fichier *demo_bt500.py*.

demo_bt500.py:

```
import argparse
import csv
import sys
import pprint

import numpy as np
from scipy import linalg


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

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

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

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

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

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

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

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

    return data3d


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

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

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


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

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


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

```
        axis=1)
    s_jk_std = 1.0 / np.sqrt(np.maximum(0., den))
    return s_jk_std


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

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

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

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

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

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

    else:
        NotImplementedError

    return u



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

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

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

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

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

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

    itr = 0
    while True:

        u_jk_prev = u_jk
```

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

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

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

        itr += 1

        delta_u_jk = linalg.norm(u_jk_prev - u_jk)

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

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

        if delta_u_jk < DELTA_THR:
            break

        if itr >= MAX_ITR:
            break

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

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

    return {
        'mos_j': list(u_jk),
        'sos_j': list(u_jk_std),
        'bias_i': list(b_i),
        'inconsistency_i': list(sig_i),
    }


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

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

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

```
o_jir = read_csv_into_3darray(input_csv)

ret = run_alternating_projection(o_jir)

pprint.pprint(ret)
```

Annexe 2
de la Partie 1

Description d'un format commun pour l'échange de fichier

L'utilisation d'un format commun pour l'échange de fichier vise à faciliter l'échange de données entre des laboratoires participant collectivement à une campagne internationale d'évaluations subjectives.

Toutes les évaluations subjectives s'articulent autour des cinq phases successives interdépendantes suivantes: préparation du test, exécution du test, traitement des données, présentation et interprétation des résultats. Généralement, pour des campagnes internationales de grande envergure, les tâches sont réparties entre les différents laboratoires participants:

– un laboratoire est chargé d'organiser le test en coopération avec les autres parties, en identifiant les paramètres de qualité à évaluer, les séquences d'images de test à utiliser (dont le contenu est généralement «critique» mais pas excessivement), la structure du test (méthodologie, distances d'observation, organisation de la séance, ordre de présentations des sujets du test, par exemple) et les conditions du test (conditions d'observation, discours d'introduction, par exemple);
– les laboratoires volontaires fourniront les séquences d'images de test traitées conformément aux techniques appropriées représentatives du paramètre de qualité à évaluer (par simulation ou à l'aide d'équipement matériel);
– un autre partenaire est chargé du montage de la bande d'essai;
– des laboratoires volontaires différents font le test en utilisant la bande préalablement montée. Il peut s'agir d'un test aveugle. Dans ce cas, le laboratoire exécutera le test en regroupant les notes attribuées par les observateurs sans nécessairement connaître les paramètres de qualité à évaluer;
– généralement, un autre participant coordonnera la collecte des données brutes résultantes en vue de leur traitement et de l'édition des résultats, ce qui peut également être effectué de façon aveugle;
– les résultats sont enfin interprétés, à partir d'un texte d'un tableau ou d'un graphique; puis un rapport final est publié.

Le format proposé permet de regrouper les résultats remis conformément aux procédures de test définies pendant la phase de définition du test.

Le format est conforme aux méthodes d'évaluation décrites dans la Partie 1 et la Partie 2 de la présente Recommandation.

Il se compose de fichiers de texte, dont la structure est illustrée dans les Tableaux 1-4 et 1-5. Sa syntaxe est structurée en étiquettes et champs auxquels s'ajoute un ensemble limité de symboles réservés («[», «]», « », «¿» et «=», par exemple).

Il n'y a aucune limitation quant à la capacité (nombre de laboratoires participants, observateurs, séquences de test, paramètres de qualité, limites des échelles de notation ou type de périphérique utilisé pour les notations, par exemple).

TABLEAU 1-4

Format de fichier de données pour l'identification des résultats

Format et syntaxe du fichier d'identification	Commentaires
[Structure du test]↵ Type = *«DSCQS» ou «DSIS I», «DSIS II», etc.*↵ Nombre de séances = *1 ≤ entier ≤ x*↵ Minimum de l'échelle= *entier*↵	[Identificateur de section] Identification de la méthodologie Rec. UIT-R BT.500 utilisée Nombre de séances(1) par test Définition de l'échelle (voir les spécifications propres à la méthodologie, s'il y en a)
Maximum de l'échelle = *entier*↵ Taille de l'écran = *entier*↵	Longueur de la diagonale (pouces)
Fabricant et modèle de l'écran = *chaîne de caractères*↵	
[RÉSULTATS] ↵ Nombre de résultats = *1 ≤ entier ≤ y*↵ Résultat(*j*).Nom de fichier(s) = *chaîne de caractèreres*.DAT↵	[Identificateur de section] Nombre de fichiers résultats(1) pris en considération Nom de fichier complet.DAT (voir le Tableau 1-5), y compris le chemin
.... Résultat(*j*).Nom = *chaîne de caractères* ↵ Résultat(*j*).Laboratoire = *chaîne de caractères* ↵ Résultat(*j*).Nombre d'observateurs = *1 ≤ entier ≤ N*↵ Résultat(*j*).Initialisation = *«Oui» ou «Non»* ↵	Nom du fichier résultats habituel Identification du laboratoire effectuant le test Nombre total d'observateurs Indique si les notes recueillies pendant l'initialisation sont incluses dans le fichier DAT joint
[Résultat(*j*).Séance(*i*).Observateurs] ↵ O(*k*).Nom = *chaîne de caractères*↵ O(*k*).Prénom = *chaîne de caractères*↵ O(*k*).Sexe = *«F» ou «M»* ↵ O(*k*).Age = *entier*↵ O(*k*). Activité professionnelle = *chaîne de caractères*↵ O(*k*).Distance = *entier*↵	[Identificateur de section] Identification de l'observateur Facultatif Facultatif Principaux groupes socio-économiques (ouvriers, étudiants, par exemple) Distance d'observation, en hauteur d'écran (3 H, 4 H, 6 H, par exemple)

(1) Séance: un test peut être divisé en un certain nombre de séances différentes pour respecter les impératifs en ce qui concerne la durée maximale du test. Les mêmes observateurs ou des observateurs différents peuvent assister à différentes séances durant lesquelles il leur sera demandé d'évaluer différentes configurations. Le regroupement des notes recueillies au cours de différentes séances donne un ensemble détaillé de résultats de test (nombre de présentations multiplié par le nombre de notes par présentation). Les résultats peuvent être joints dans différents. DAT qui seront remis pour chaque exécution.

TABLEAU 1-5

Format de fichier de texte des résultats.DAT des données brutes

Format et syntaxe du fichier nom_de_fichier.DAT	Commentaires
entier entier entier.....↵	Un fichier de données brutes DAT se compose de valeurs de notation séparées par un espace.
entier entier entier.....↵	On utilisera une ligne par observateur.
entier entier entier.....↵	Les données brutes sont mémorisées dans leur ordre d'entrée.
.....	Les données peuvent être réparties entre différents fichiers DAT identifiés dans le Tableau 1-4 par Résultat(j). Fichier(s)(1).

(1) Voir la Note(1) du Tableau 1-4.

Annexe 3
(pour information)
de la Partie 1

Caractéristiques de dégradation du contenu de l'image

A3-1 Introduction

Une fois mis en œuvre, un système devra traiter une gamme potentiellement étendue de programmes et il risque d'infliger à certains d'eux une perte de qualité. Pour voir si un système convient, il faut connaître la proportion des programmes qui lui causent des difficultés et la perte de qualité qui en résulte. On a en effet besoin, pour le système considéré, d'une caractéristique de défaillance fonction du contenu de l'image.

Cette caractéristique est particulièrement importante pour les systèmes dont la qualité ne se dégrade pas progressivement quand l'image devient de plus en plus critique. Par exemple, certains systèmes numériques et adaptatifs peuvent conserver une haute qualité sur toute une large gamme de types de programmes mais se détériorer au-delà.

A3-2 Établissement de la caractéristique de dégradation

Le concept de caractéristique fonction du contenu de l'image définit la proportion des programmes susceptibles de se présenter à long terme et pour lesquels le système donne un certain niveau de qualité. C'est ce qu'illustre la Fig. 1-5.

On peut obtenir, en quatre étapes, la caractéristique de dégradation du contenu de l'image:

– *Étape 1*: détermine une mesure algorithmique de la «criticité» qui doit pouvoir permettre de classer par ordre de mérite un certain nombre de séquences d'images auxquelles le système ou la catégorie de systèmes concernés ont infligé des distorsions, de sorte que le classement correspond à celui qu'auraient donné des observateurs humains chargés de cette tâche. La mesure de criticité peut prendre en compte une modélisation de la vision.

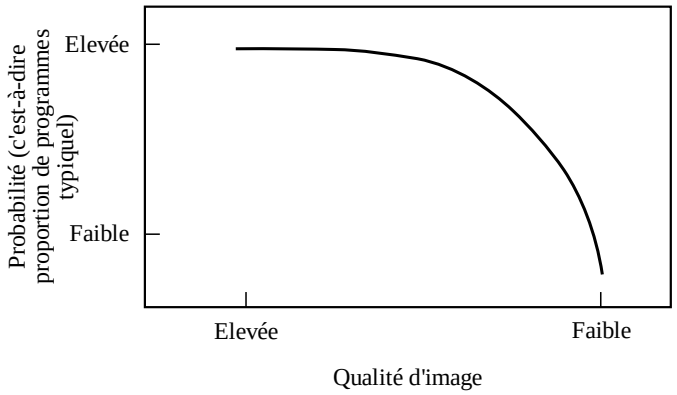
– *Étape 2*: établit, en appliquant la mesure de criticité à un grand nombre d'échantillons de programmes types de télévision, une distribution qui estime la probabilité que se présentent des images de niveaux de criticité divers pour le système ou les catégories de systèmes considérés. La Figure 1-6 présente un exemple de ce genre de distribution.

– *Étape 3*: établit, de façon empirique, la faculté du système de conserver la qualité quand le niveau de criticité du programme augmente. Dans la pratique, il faut évaluer subjectivement la qualité que donne le système avec les programmes choisis pour échantillonner la gamme de criticité définie à l'Étape 2. Il en résulte une fonction qui met en relation la qualité donnée par le système et le niveau critique du programme. La Figure 1-7 donne un exemple de cette fonction.

– *Étape 4*: combine les résultats des Étapes 2 et 3 pour établir une caractéristique de dégradation du contenu de l'image, comme celle de la Fig. 1-5.

FIGURE 1-5

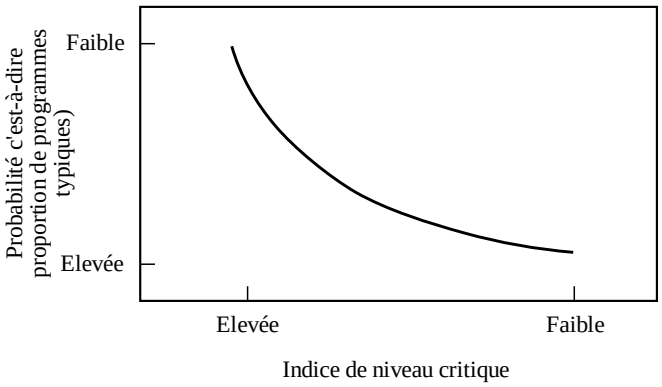
Représentation graphique d'un exemple de caractéristiques de défaillance fonction du contenu de l'image



BT.0500-01-5

FIGURE 1-6

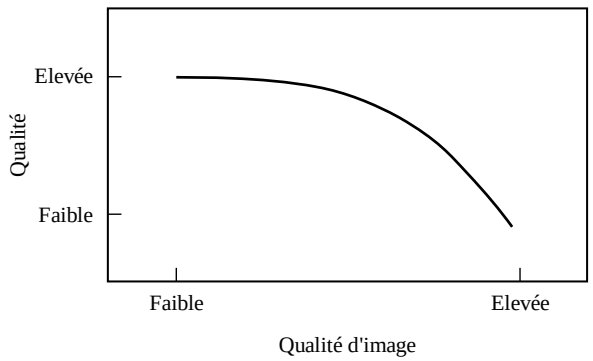
Probabilité que se présentent des images d'un certain niveau critique



BT.0500-01-6

FIGURE 1-7

Exemple de courbe de la qualité en fonction du niveau critique du programme



BT.0500-01-7

A3-3 Utilisation de la caractéristique de dégradation

La caractéristique de dégradation, qui donne une idée globale de la qualité à espérer pour tous les types de programmes possibles, est un moyen essentiel pour étudier si un système convient. On peut s'en servir de trois façons:

– pour optimiser les caractéristiques d'un système (par exemple, la resolution à la source, le débit binaire, la largeur de bande) au moment de sa conception afin de l'adapter au mieux aux exigences d'un service;

– pour étudier si un système donné conviendra (c'est-à-dire prévoir la conséquence et la gravité des défaillances pendant l'exploitation);

– pour voir, parmi plusieurs systèmes, ceux qui conviennent le mieux (c'est-à-dire comparer les caractéristiques de dégradation et déterminer quel système conviendra le mieux à l'usage envisagé). On notera que, si plusieurs systèmes possibles de type semblable peuvent avoir le même indice de criticité, des systèmes de types différents peuvent en avoir de distincts. Toutefois, bien que la caractéristique de dégradation n'exprime que la probabilité d'observer dans la pratique différents niveaux de qualité, on peut comparer directement les caractéristiques même si elles résultent d'indices de criticité distincts et propres au système.

Alors que la méthode décrite ci-dessus donne une façon de mesurer la caractéristique de dégradation du contenu de l'image d'un système, il n'est pas sûr qu'elle puisse prévoir si un système sera acceptable pour un téléspectateur. On obtient cette information en faisant observer, par un certain nombre d'observateurs, des programmes codés selon le système en question et en examinant leurs commentaires.

Un exemple de caractéristiques de dégradation du contenu de l'image pour la télévision numérique est donné dans l'Annexe 1 de la Partie 3.

Annexe 4
(pour information)
de la Partie 1

Méthode de détermination d'une caractéristique de dégradation composite en fonction du contenu du programme et des conditions de transmission

A4-1 Introduction

Une caractéristique de dégradation composite établit une relation entre la qualité de l'image perçue et la probabilité pratique de l'obtenir, en considérant explicitement le contenu du programme et les conditions de transmission.

On pourrait obtenir une telle caractéristique au moyen d'études subjectives avec un nombre suffisant d'observations, d'essais et de points de réception pour avoir un échantillon représentatif de la population des contenus de programmes possibles et des conditions de transmission. Dans la pratique, toutefois, une expérience semblable risque d'être irréalisable.

La présente Annexe décrit une autre méthode, plus facile à mettre en œuvre, permettant de déterminer la caractéristique de dégradation composite. Cette méthode comprend trois étapes:

– analyse du contenu du programme;

– analyse du canal de transmission;

– établissement des dégradation composite.

A4-2 Analyse du contenu du programme

Cette étape comprend deux opérations. On définit d'abord une mesure appropriée du contenu du programme. Puis on évalue la façon dont se répartissent les probabilités des résultats de la mesure.

Une mesure du contenu du programme est une statistique qui révèle les aspects du contenu du programme, lesquels soulignent la faculté du ou des systèmes considérés à donner une reproduction du programme perçue comme étant fidèle. Il serait évidemment avantageux que cette mesure soit fondée sur un modèle de perception approprié. Toutefois, en l'absence d'un tel modèle, il peut suffire d'avoir une mesure qui rende compte de certains aspects de l'importance de la diversité spatiale, dans les trames ou images vidéo et entre elles, pourvu qu'elle présente une relation à peu près uniforme avec la qualité perçue de l'image. Il peut être nécessaire de recourir à des modes de mesure différents pour des systèmes (ou catégories de systèmes) qui ont des méthodes de représentation de l'image complètement différentes.

Une fois qu'un mode de mesure approprié a été choisi, il faut estimer avec quelle probabilité les valeurs statistiques possibles surviennent. Cela peut se faire de deux façons différentes:

– avec la méthode empirique, on analyse un échantillon aléatoire de, par exemple, 200 segments de programme de 10 s, à un format de production qui, du point de vue de la résolution, de la fréquence d'image et du format d'image, convienne au(x) système(s) considéré(s). L'analyse de ces échantillons fournit les fréquences relatives d'apparition des valeurs statistiques que l'on prend comme estimations de la probabilité d'apparition dans la pratique; ou

– avec la méthode théorique, on estime les probabilités au moyen d'un modèle théorique. On notera que, bien que la méthode empirique soit préférée, il peut être nécessaire, dans certains cas, de recourir à la méthode théorique (par exemple, lorsqu'on n'a pas assez de renseignements sur le contenu du programme, notamment lorsque de nouvelles technologies de production apparaissent).

Les analyses ci-dessus aboutiront, pour les valeurs statistiques du contenu, à une distribution de probabilité (voir aussi l'Annexe 3). On les combinera avec les résultats de l'analyse des conditions de transmission pour préparer l'étape finale de la méthode.

A4-3 Analyse du canal de transmission

Cette étape comprend aussi deux opérations. On définit d'abord une mesure de la qualité du canal de transmission. Puis on évalue la façon dont se répartissent les probabilités des résultats de la mesure.

Une mesure du canal de transmission est une statistique qui révèle les aspects de la qualité du canal, lesquels influencent la faculté du ou des systèmes considérés de donner une reproduction du programme perçue comme étant fidèle. Il serait évidemment avantageux que cette mesure soit fondée sur un modèle de perception approprié. Toutefois, en l'absence d'un tel modèle, il peut suffire d'avoir une mesure qui rende compte de certains aspects des contraintes qu'impose le canal, pourvu qu'elle présente une relation à peu près uniforme avec la qualité perçue de l'image. Il peut être nécessaire de recourir à des modes de mesure différents pour des systèmes (ou catégories de systèmes) qui ont des méthodes de codage de canal complètement différentes.

Une fois qu'un mode de mesure approprié a été choisi, il faut estimer avec quelle probabilité les valeurs statistiques possibles surviennent. Cela peut se faire de deux façons différentes:

– avec la méthode empirique, la qualité de la voie est mesurée par exemple en 200 instants et points de réception choisis au hasard. L'analyse de ces échantillons fournit les fréquences

relatives d'apparition des valeurs statistiques que l'on prend comme estimations de la probabilité d'apparition dans la pratique; ou

– avec la méthode théorique, on estime les probabilités au moyen d'un modèle théorique. On notera que, bien que la méthode empirique soit préférée, il peut être nécessaire, dans certains cas, de recourir à la méthode théorique (par exemple, lorsqu'on n'a pas assez de renseignements sur la qualité de la voie, notamment lorsque de nouvelles technologies de transmission apparaissent).

Les analyses ci-dessus aboutiront, pour les valeurs statistiques du canal, à une distribution de probabilité. On les combinera avec les résultats de l'analyse du contenu du programme pour préparer l'étape finale de la méthode.

A4-4 Établissement des caractéristiques de dégradation composite

À cette étape, on procède à une expérimentation subjective au cours de laquelle le contenu du programme et les conditions de transmission varient tous deux selon les probabilités calculées au cours des deux premières étapes.

La méthode fondamentale mise en œuvre ici est la procédure à double stimulus utilisant une échelle de qualité continue et notamment la version à 10 s recommandée pour les séquences animées (voir l'Annexe 2 de la Partie 2). La référence y est une image de qualité studio au format approprié (par exemple, avec la résolution, la fréquence et le format d'image convenant au(x) système(s) considéré(s)). En revanche, au cours de l'essai, on présente la même image que celle qui serait reçue avec le ou les systèmes considérés dans des conditions de transmission choisies.

Le matériel d'essai et les conditions de transmission sont choisis d'après les probabilités établies au cours des deux premières étapes de la méthode. Parmi les segments de matériel d'essai, qui ont chacun été étudiés en vue de déterminer leur valeur essentielle selon la statistique du contenu, on a un ensemble de sélection. On prélève dans cet ensemble du matériel de façon qu'il couvre toute la gamme possible des valeurs statistiques et on en prend d'autant plus que le niveau est plus critique. On choisit de même les valeurs statistiques possibles pour le canal. Ensuite, ces deux causes de variation d'origine indépendante sont combinées de façon aléatoire pour former une combinaison de contenus et de conditions de transmission de probabilité donnée.

Les résultats de ces études, qui établissent une relation entre la qualité d'image perçue et sa probabilité d'apparition dans la pratique, servent ensuite à estimer si un système convient ou à comparer des systèmes selon qu'ils conviennent plus ou moins bien.

Annexe 5
(pour information)
de la Partie 1

Effet contextuel

Il se produit des effets contextuels lorsque l'évaluation subjective d'une image est influencée par l'ordre et par la gravité des dégradations présentées. Par exemple, si une image très dégradée est présentée après une séquence d'images peu dégradées, les observateurs peuvent, par inadvertance, évaluer cette image à un niveau inférieur que celui où ils l'auraient peut-être située normalement.

Quatre laboratoires, opérant dans des pays différents, ont analysé les effets contextuels possibles associés aux résultats fournis par trois méthodes d'évaluation de la qualité des images (la méthode

DSCQS; la variante II de la méthode DSIS; et une méthode de comparaison). Le matériel d'essai était produit par codage MPEG (ML@MP) avec réduction de la résolution horizontale. Dans chaque série d'essais, on appliquait quatre conditions d'essai fondamentales (B1, B2, B3, B4) et six conditions d'essai contextuelles; l'une des séries décrivait de faibles dégradations contextuelles, l'autre de fortes dégradations. Les trois méthodes d'essai étaient appliquées aux deux séries. Les effets contextuels représentent la différence entre les résultats de l'essai avec prédominance de dégradations faibles et ceux de l'essai avec prédominance de dégradations fortes. Les effets contextuels ont été déterminés dans le cas des conditions d'essai fondamentales B2 et B3.

Les résultats obtenus, tous laboratoires confondus, ne font pas apparaître d'effets contextuels pour la méthode DSCQS. Ces effets ont été clairement mis en évidence dans le cas de la méthode DSIS et de la méthode de comparaison; l'effet le plus fort a été obtenu dans la variante II de la méthode DSIS. Les résultats montrent que les essais avec prédominance de dégradations faibles peuvent conduire à une sous-évaluation de l'image, alors que les essais avec prédominance de dégradations fortes peuvent donner une surévaluation.

Les résultats de cette étude donnent à penser que la méthode DSCQS est la méthode la plus efficace pour réduire à un minimum les effets contextuels aux fins de l'évaluation subjective de la qualité des images recommandée par l'UIT-R.

Le Rapport UIT-R BT.1082 donne de plus amples renseignements sur l'étude décrite ci-dessus.

Annexe 6
(pour informations)
de la Partie 1

Informations spatiales et spatio-temporelles

Les informations spatio-temporelles sont indiquées ci-après sous la forme de valeurs uniques pour chaque trame dans une séquence d'essai complète. Cela se traduit par une série temporelle de valeurs qui varieront généralement un peu. Les valeurs d'information de perception indiquées ci-dessous éliminent cette variabilité au moyen d'une fonction à maximum (valeur maximale pour la séquence). La variabilité proprement dite peut servir à d'utiles études, par exemple pour des tracés d'informations spatio-temporelles trame par trame. Le recours à des distributions des informations pour l'ensemble d'une séquence d'essai permet aussi de mieux évaluer les scènes comportant des coupes.

Information de perception spatiale (SI): grandeur représentative du degré de détail spatial d'une image. Cette valeur augmente généralement avec la complexité spatiale des scènes. Cette grandeur n'est pas censée mesurer l'entropie ni être associée aux informations définies en théorie de la communication. L'information de perception spatiale, *SI*, est fondée sur le filtre de Sobel. Chaque trame vidéo (plan de la luminance) à l'instant n (F_n) est d'abord filtrée par le filtre de Sobel [Sobel(F_n)]. On calcule ensuite, pour chaque trame passant par le filtre de Sobel, l'écart type pour l'ensemble des pixels (std_{space}). Cette opération est répétée pour chaque trame de la séquence vidéo et permet d'obtenir une série temporelle d'informations spatiales sur la scène. La valeur maximale contenue dans la série temporelle ($\max_{time}$) est choisie pour représenter le contenu de la scène en informations spatiales. Ce processus peut être représenté sous forme d'équation, comme suit:

$$SI = \max_{time} \{std_{space} [\text{Sobel}(F_n)]\}$$

Information de perception temporelle (TI): grandeur représentative du degré de changements temporels d'une séquence vidéo. Cette valeur augmente généralement avec l'animation des séquences. Cette grandeur n'est pas censée mesurer l'entropie ni être associée aux informations définies en théorie de la communication.

On calcule la grandeur d'information temporelle, *TI*, en tant que valeur maximale dans le temps ($\max_{time}$) de l'écart type pour l'ensemble de l'espace (std_{space}) de $M_n(i, j)$ pour tous les *i* et tous les *j*, soit:

$$TI = \max_{time}\{std_{space}[M_n(i, j)]\}$$

où $M_n(i, j)$ est la différence entre les pixels au même point dans la trame mais appartenant à deux trames successives, soit:

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

où $F_n(i, j)$ est le pixel situé dans la *i*ème ligne et dans la *j*ème colonne de la *n*ème trame dans le temps.

NOTE – Pour les scènes comportant des coupes, deux valeurs peuvent être indiquées: l'une où la coupe est incluse dans la grandeur d'information temporelle, l'autre où elle en est exclue.

Annexe 7
(pour information)
de la Partie 1

Termes et définitions

Algorithme	une ou plusieurs opérations de traitement d'image
AVI	entrelacement audio vidéo (*audio video interleaved*)
CCD	dispositif à transfert de charge (*charge coupled device*)
CI	intervalle de confiance (*confidence interval*)
CIF	format intermédiaire commun (*common intermediate format*) (défini dans la Recommandation UIT-T H.261 pour des visiophones: 352 lignes × 288 pixels)
CRT	tube cathodique (*cathode ray tube*)
DSCQS	méthode à double stimulus utilisant une échelle de qualité continue (*double stimulus using a continuous quality scale method*)
DSIS	méthode à double stimulus utilisant une échelle de dégradation (*double stimulus using an impairment scale method*)
LCD	affichage à cristaux liquides (*liquid crystal display*)
MOS	note moyenne d'opinion (*mean opinion score*)
PDP	écran à plasma (*plasma display panel*)
PS	segment de programme (*programme segment*)
QCIF	quart de CIF (*quarter CIF*) (défini dans la Recommandation UIT-T H.261 pour des visiophones: 176 lignes × 144 pixels)
S/N	rapport signal sur bruit (*signal-to-noise ratio*)

SAMVIQ	évaluation subjective de la qualité vidéo multimédia (*subjective assessment of multimedia video quality*)
SC	méthode de comparaison de stimulus (*stimulus comparison method*)
Scène	contenu audiovisuel
sdt	écart type (*standard deviation*)
Séquence	scène avec traitement combiné ou sans traitement
SI	information spatiale (*spatial information*)
SIF	format intermédiaire standard (*standard intermediate format*) [défini dans la norme ISO 11172 (MPEG-1): 352 lignes × 288 pixels × 25 images/s et 352 lignes × 240 pixels × 30 images/s]
SP	présentation simultanée (*simultaneous presentation*)
SQCIF	sub-QCIF
SS	méthode à un seul stimulus (*single stimulus method*)
SSCQE	méthode d'évaluation continue de la qualité avec stimulus unique (*single stimulus using a continuous quality evaluation method*)
TI	information temporelle (*temporal information*)
TP	présentation de l'essai (*test presentation*)
TS	séance d'essai (*test session*)
VTR	magnétoscope (*video tape recorder*)

PARTIE 2

Description des méthodologies d'évaluation subjective des images

1 Introduction

On trouvera ci-après une présentation détaillée de chaque méthodologie d'évaluation des images nécessaire pour effectuer des évaluations subjectives de la qualité des images. Dans certains cas, les caractéristiques varient par rapport aux caractéristiques communes d'évaluation données au § 2 de la Partie 1.

Afin de veiller à ce que les résultats des évaluations subjectives de la qualité des images puissent être interprétés correctement par d'autres laboratoires, il est important que des notes détaillées concernant les procédures soient disponibles et que toute variation par rapport à la méthodologie utilisée soit consignée avec toutes les informations supplémentaires dont aurait besoin un autre laboratoire souhaitant répéter la procédure d'évaluation.

2 Méthodologies d'évaluation des images recommandées

Annexe 1	Méthode à double stimulus utilisant une échelle de dégradation (DSIS)
Annexe 2	Méthode à double stimulus utilisant une échelle de qualité continue (DSCQS)
Annexe 3	Méthode à un seul stimulus (SS)
Annexe 4	Méthode de comparaison des stimulus
Annexe 5	Méthode d'évaluation continue de la qualité avec stimulus unique (SSCQE)
Annexe 6	Méthode d'évaluation continue à double stimulus simultanés (SDSCE)
Annexe 7	Méthode d'évaluation subjective de la qualité vidéo multimédia (SAMVIQ)
Annexe 8	Protocole d'observation par des spécialistes (EVP) pour l'évaluation de la qualité du matériel vidéo

3 Remarques

D'autres techniques comme les méthodes à échelle multidimensionnelle et à variables aléatoires multiples sont décrites par le Rapport UIT-R BT.1082 et sont encore à l'étude.

Toutes les méthodes décrites jusqu'ici présentent des avantages et des inconvénients; il n'est pas encore possible d'en recommander absolument une plutôt qu'une autre. Il incombe donc au chercheur de choisir la méthode qui convient le mieux aux conditions qui prévalent.

Les limites inhérentes aux différentes méthodes donnent à penser qu'il pourrait être déraisonnable de trop insister sur une seule méthode. Il peut donc sembler plus judicieux d'envisager des approches plus «complètes», c'est-à-dire d'utiliser plusieurs méthodes, ou une approche multidimensionnelle.

Annexe 1
de la Partie 2

Méthode à double stimulus utilisant une échelle de dégradation (DSIS) (méthode UER)

A1-1 Description générale

Dans le cadre d'une évaluation typique, on peut chercher à évaluer soit un nouveau système, soit l'effet d'une dégradation due à la transmission. La personne chargée d'organiser l'évaluation devra tout d'abord choisir un matériel d'essai suffisant pour que l'évaluation puisse être significative et définir les conditions d'observations à utiliser. Si la variation de paramètres présente de l'intérêt, il est nécessaire de choisir un jeu de valeurs de paramètres réparties sur l'échelle de dégradation à intervalles plus ou moins égaux. Si on évalue un nouveau système dont on ne peut faire varier de cette façon les valeurs des paramètres, il faut alors, soit ajouter de nouvelles dégradations subjectivement identiques, soit utiliser une autre méthode, par exemple celle indiquée dans l'Annexe 2 de la Partie 2.

La méthode à double stimulus utilisation une échelle de dégradation (DSIS) (méthode UER) est cyclique en ce sens que l'on présente d'abord une image de référence non dégradée puis la même image dégradée à l'observateur qui est ensuite prié de donner son avis sur la seconde image, tout en gardant à l'esprit la première. Au cours des séances, qui durent au plus une demi-heure, une série d'images ou de séquences d'images, couvrant toutes les combinaisons requises, sont présentées à l'observateur. Les ordres de présentation des images et des dégradations sont aléatoires. L'image de référence fait partie des images ou des séquences d'images à évaluer. À l'issue de la série de séances, on calcule la note moyenne pour chaque condition d'essai et chaque image d'essai.

Cette méthode fait appel à l'échelle de dégradation qui donne généralement des résultats plus stables pour des dégradations faibles que pour des dégradations importantes. Bien qu'elle ait parfois été utilisée avec des gammes de dégradations limitées, cette méthode convient mieux pour une gamme de dégradations complète.

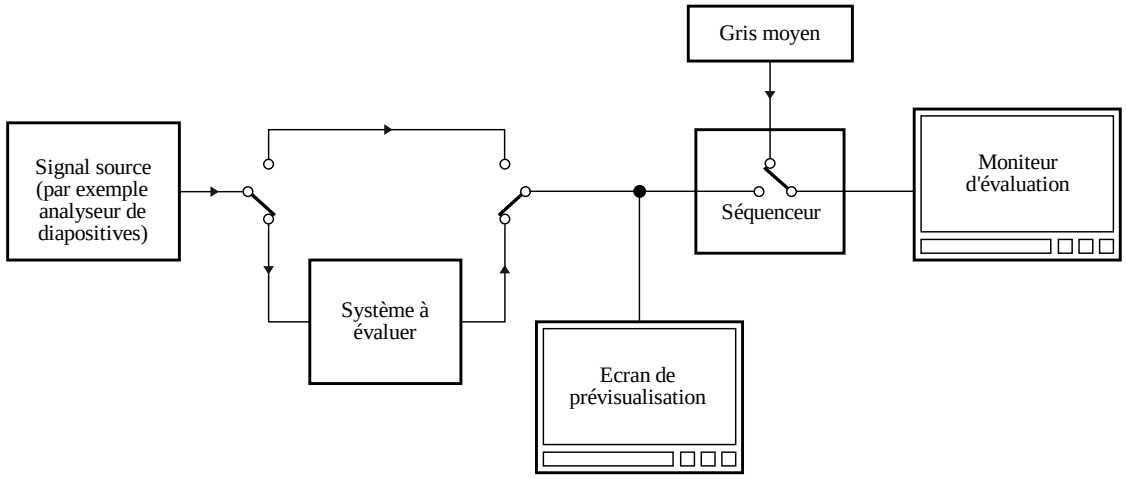
A1-2 Mode opératoire général

Désigne la façon de définir ou de choisir, conformément au § 2 de la Partie 1, les conditions d'observation, les signaux source, le matériel d'essai, les observateurs et la présentation des résultats.

Le système d'essai aura la configuration indiquée à la Fig. 2-1.

FIGURE 2-1

Configuration du système d'essai pour la méthode DSIS



On présente aux observateurs un écran d'évaluation qui reçoit un signal au moyen d'un séquenceur. Le trajet du signal jusqu'au séquenceur peut être, soit direct (le signal vient de la source), soit indirect (le signal passe par le système à évaluer). Les observateurs voient défiler devant eux une série d'images ou de séquences d'images d'essai. Ces images sont présentées par paire: la première image de la paire provient directement de la source, la seconde est la même image qui est passée par le système à évaluer.

A1-3 Présentation du matériel d'essai

Une séance d'évaluation se compose d'un certain nombre de présentations. Il existe deux variantes (I et II) de la structure des présentations.

Variante I: Image ou la séquence de référence et l'image ou la séquence d'essai ne sont présentées qu'une seule fois (Fig. 2-2a)).

Variante II: Image ou la séquence de référence et l'image ou la séquence d'essai sont présentées deux fois (Fig. 2-2b)).

La variante II, plus longue que la variante I, peut être utilisée si une discrimination de très petites dégradations est nécessaire ou si des séquences animées sont soumises à des essais.

A1-4 Échelles d'évaluation

Il convient d'utiliser une échelle de dégradation à cinq notes:

5 imperceptible

4 perceptible mais non gênant

3 légèrement gênant

2 gênant

1 très gênant.

Les observateurs utiliseront un formulaire représentant très clairement l'échelle, avec des cases numérotées ou un autre moyen pour consigner les notes.

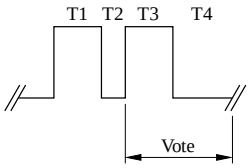
A1-5 Introduction aux évaluations

Au début de chaque séance, une explication est donnée aux observateurs sur le type d'évaluation, l'échelle d'évaluation, la séquence elle-même et le séquencement (image de référence, gris, image d'essai, vote). Il convient que la dynamique et le type de dégradations à évaluer soient illustrés sur des images autres que celles utilisées dans les essais mais ayant une sensibilité comparable. Il ne faut pas en déduire que la plus mauvaise qualité observée correspond nécessairement à la notation subjective la plus basse. Les observateurs seront priés de fonder leur jugement sur l'impression globale donnée par l'image et d'exprimer ce jugement à l'aide des termes utilisés pour définir l'échelle d'évaluation subjective.

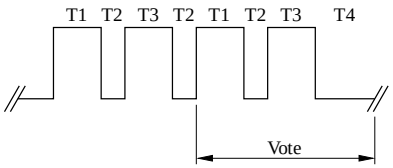
Les observateurs seront priés de regarder l'image pendant toute la durée de T1 et T3. Le vote sera autorisé uniquement pendant T4.

FIGURE 2-2

Structure de la présentation du matériel d'essai



a) Variante I



b) Variante II

BT.0500-02-2

Phases de la présentation:
T1 = 10 s Image de référence
T2 = 3 s Gris moyen produit par un niveau vidéo d'environ 200 mV
T3 = 10 s Condition à l'essai
T4 = 5-11 s Gris moyen
L'expérience montre que le fait de prolonger les phases T1 et T3 au-delà de 10 s ne permet pas aux observateurs de mieux évaluer les images ou les séquences.

A1-6 Séance de test

Les images et les dégradations seront présentées dans un ordre pseudo-aléatoire et, de préférence, dans un ordre différent pour chaque séance. En tout état de cause, la même image ne devra jamais être présentée deux fois consécutivement, que les niveaux de dégradation soient les mêmes ou qu'ils soient différents.

La gamme des dégradations devra être choisie de telle sorte que toutes les notes soient utilisées par la majorité des observateurs; l'objectif est de tendre vers une moyenne générale (moyenne de tous les jugements émis au cours de l'expérience) voisine de 3.

Une séance ne dépassera pas une trentaine de minutes, y compris les explications et présentations préliminaires; la séquence d'essai pourra commencer par quelques images représentatives de la gamme des dégradations; les jugements relatifs à ces images ne seront pas pris en considération dans les résultats finals.

L'Annexe 2 de la Partie 1 donne des indications supplémentaires pour le choix des niveaux de dégradation.

Annexe 2
de la Partie 2

Méthode à double stimulus utilisant une échelle de qualité continue (DSCQS)

A2-1 Description générale

Dans une évaluation typique, on peut évaluer soit un nouveau système, soit les effets de la transmission sur la qualité. On estime que la méthode à double stimulus est particulièrement utile lorsqu'il n'est pas possible de créer des conditions expérimentales et des stimulus d'essai représentant toute la gamme de qualité.

La méthode est cyclique en ce sens que l'on présente à l'observateur une paire d'images, chacune provenant de la même source, l'une ayant passé par le système à évaluer et l'autre venant directement de la source. L'observateur est prié d'évaluer la qualité des deux images.

Au cours des séances qui durent au plus 30 min, on présente à l'observateur une série de paires d'images, les images constituant la paire alternant aléatoirement. Les images et les dégradations, couvrant toutes les combinaisons requises, sont également présentées dans un ordre aléatoire. À l'issue des séances, on calcule les moyennes pour chaque condition expérimentale et chaque image d'essai.

A2-2 Mode opératoire général

Désigne la façon de définir ou de choisir, conformément au § 2 de la Partie 1, les conditions d'observation, les signaux source, le matériel d'essai, les observateurs et la présentation des résultats. La séance d'essai est décrite au § A1-6 de l'Annexe 1 de la Partie 2.

Le système d'essai aura la configuration indiquée à la Fig. 2-3.

A2-3 Présentation du matériel d'essai

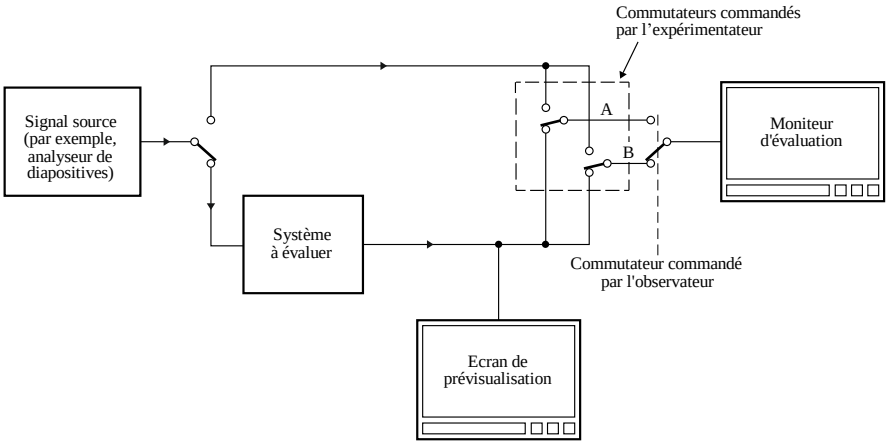
Une séance d'évaluation comprend un certain nombre de présentations. Dans le cas de la variante I, qui ne nécessite qu'un seul observateur, l'observateur peut, pour chaque présentation, passer du trajet A au trajet B et inversement jusqu'à ce que l'observateur ait mentalement la mesure de la qualité associée à chaque signal. Il peut répéter cette opération deux ou trois fois pendant des laps de temps ne dépassant pas 10 s. Dans la variante II, qui fait appel à plusieurs observateurs simultanément, avant d'enregistrer les résultats, chaque paire de conditions est présentée une ou plusieurs fois pendant un laps de temps égal, afin de permettre aux observateurs de mesurer mentalement les qualités associées à ces conditions. Ensuite, chaque paire est visualisée une ou plusieurs fois tandis que les résultats sont enregistrés. Le nombre de répétitions dépend de la longueur des séquences d'essai. Pour des images fixes, une séquence de 3 à 4 s et cinq répétitions (avec notation pendant les deux dernières) peut

convenir. Pour des images en mouvement avec des défauts variant dans le temps, une séquence de 10 s avec deux présentations (et notation pendant la seconde) peut être appropriée. La structure des présentations est illustrée à la Fig. 2-4.

Si des considérations pratiques limitent la durée des séquences disponibles à moins de 10 s, il est possible de recourir à des compositions utilisant ces séquences plus courtes sous forme de segments afin d'étendre jusqu'à 10 s le temps de visualisation. Pour réduire au minimum la discontinuité aux jonctions, des segments de séquence successifs peuvent être inversés dans le temps (appelés parfois visualisation «palindromique»). Il convient cependant de s'assurer que les conditions d'essai au cours de la visualisation de segments en sens inverse représentent des processus de causalité, c'est-à-dire qu'ils doivent être obtenus par le passage du signal source à l'envers à travers le système en cours d'évaluation.

FIGURE 2-3

Configuration du système d'essai pour la méthode DSCQS



Il y a deux variantes de cette méthode, la variante I et la variante II, exposées ci-après.

Variante I: L'observateur, habituellement seul, est autorisé à passer de la condition A à la B et inversement jusqu'à ce qu'il se soit fait une opinion sur chacune d'elles. Les trajets A et B reçoivent l'image de référence directe ou l'image qui est passée par le système à évaluer. L'image et le trajet sont alternés de façon aléatoire d'une condition d'essai à l'autre. Ce phénomène est noté par l'expérimentateur mais non annoncé aux observateurs.

Variante II: On présente consécutivement aux observateurs les images provenant des trajets A et B, afin qu'ils se fassent une option sur chacune d'elles. Pour chaque présentation, les trajets A et B reçoivent l'image comme dans la variante I ci-dessus.

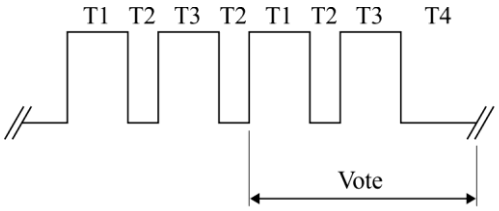
A2-4 Échelle d'évaluation

La méthode exige l'évaluation simultanée de deux versions de chaque image. Dans chaque paire d'images, l'une n'est pas dégradée alors que l'autre peut comporter ou non une dégradation. L'image non dégradée sert de référence mais les observateurs ignorent laquelle est l'image de référence. Dans la série d'essais, la position de l'image de référence est modifiée de façon pseudo-aléatoire.

Les observateurs doivent simplement évaluer la qualité globale de l'image pour chaque présentation en faisant une marque sur une échelle verticale. Les échelles verticales sont présentées par paires pour tenir compte de la double présentation de chaque image. Les échelles constituent un système de notation continu afin d'éviter les erreurs de quantification mais elles sont divisées en cinq segments égaux qui correspondent à l'échelle de qualité normale à cinq notes de l'UIT-R. Les adjectifs qui caractérisent les différents niveaux sont les mêmes que ceux utilisés normalement; dans le cas présent, ils sont indiqués comme référence et imprimés uniquement à gauche de la première échelle de chaque rangée de dix colonnes doubles sur la feuille de notation. La Figure 2-5 représente une partie d'une feuille de notation typique. On évite toute confusion possible entre les graduations de l'échelle et les résultats des essais en imprimant les échelles en bleu et en indiquant les résultats en noir.

FIGURE 2-4

Structure de la présentation du matériel d'essai



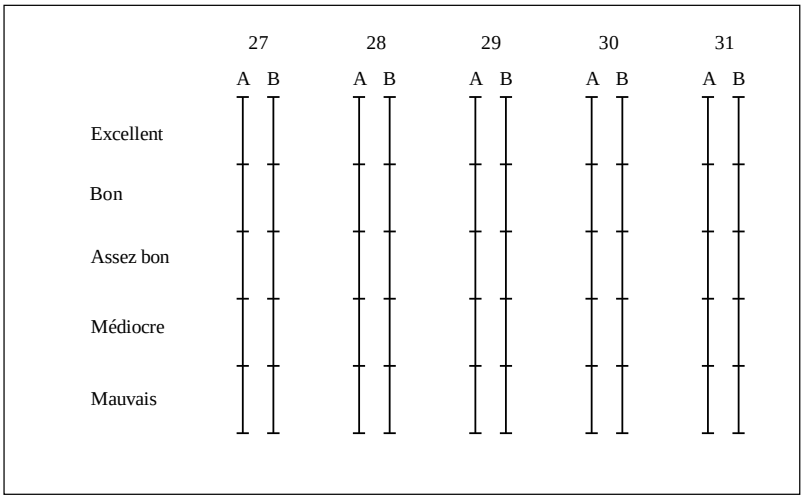
BT.0500-02-4

Phase de la présentation

T1 =	10 s	Image de référence
T2 =	3 s	Gris moyen produit par un niveau vidéo d'environ 200 mV
T3 =	10 s	Condition à l'essai
T4 =	5 to 11 s	Gris moyen

FIGURE 2-5

Partie d'un formulaire de notation de la qualité utilisant des échelles continues*



BT.0500-02-5

* Lorsqu'on planifie l'organisation des différentes présentations d'une séance d'évaluation dans le cadre de la méthode DSCQS, il est souhaitable que l'expérimentateur prévoie des contrôles donnant l'assurance que l'expérience n'est pas entachée d'erreurs systématiques. Toutefois, les modalités de ces contrôles sont à l'étude

A2-5 Analyse des résultats

Pour les deux évaluations (image de référence et image d'essai) de chaque condition d'essai, les mesures de la longueur du trait indiqué sur la feuille de notation sont converties en notes normalisées comprises entre 0 et 100, puis, les différences de notation entre l'image de référence et l'image d'essai sont calculées. Une procédure plus détaillée est décrite dans l'Annexe 2 de la Partie 1.

La pratique a montré que les notes obtenues pour différentes séquences d'essai dépendent de la criticité des séquences d'images test utilisées. Il est possible d'avoir une meilleure connaissance des caractéristiques du codec si les résultats des différentes séquences de test sont présentés séparément et non uniquement sous forme de moyennes cumulées sur l'ensemble des séquences utilisées pour l'évaluation.

Si les résultats des différentes séquences sont portés en abscisse selon un ordre fonction de la «criticité de la séquence d'essai», il est possible de présenter une description graphique brute de la caractéristique de dégradation du système à tester en fonction du contenu de l'image. Cela étant, ce type de présentation ne décrit que les caractéristiques du codec et ne donne pas d'indication sur la probabilité d'occurrence de séquences présentant un degré de criticité donné (voir l'Annexe 2 de la Partie 1). Il faut procéder à d'autres études sur la criticité des séquences d'essai et la probabilité d'occurrence de séquences présentant un degré de criticité donné avant de pouvoir avoir une idée plus complète des caractéristiques du système.

A2-6 Interprétation des résultats

Si l'on se sert de la méthode DSCQS, il pourrait être risqué, voire erroné, de tirer des conclusions sur la qualité des conditions testées en associant aux valeurs numériques obtenues avec cette méthode des adjectifs propres à d'autres protocoles d'essai (imperceptible, perceptible mais non gênant, par exemple, … adjectifs venant de la méthode DSIS).

On notera qu'il faut considérer les résultats obtenus avec la méthode DSCQS non pas comme des notes absolues mais comme des différences de notes entre l'image de référence et l'image d'essai. Il est donc incorrect d'associer aux notes un seul qualificatif même lorsqu'il s'agit des qualificatifs de cette méthode (excellent, bon, assez bon, par exemple).

Pour tout test, il est important de fixer des critères d'acceptabilité avant de commencer l'évaluation, ce qui est particulièrement vrai si l'on utilise la méthode DSCQS parce que des utilisateurs inexpérimentés ont tendance à se tromper sur la signification des valeurs de l'échelle de qualité obtenues avec cette méthode.

Annexe 3
de la Partie 2

Méthodes à un seul stimulus

Dans ce type de méthode, une seule image ou séquence d'images est présentée à l'observateur, qui fournit une notation de l'ensemble de la présentation. Le matériel d'essai peut comprendre uniquement des séquences d'essai, ou à la fois des séquences d'essai et la séquence de référence correspondante. En pareil cas, la séquence de référence est présentée sous la forme d'un stimulus indépendant pour la notation, comme n'importe quel autre stimulus d'essai.

A3-1 Mode opératoire général

Désigne la façon de définir ou de choisir, conformément au § 2 de la Partie 1, les conditions d'observation, les signaux source, la gamme des conditions et l'ancrage, les observateurs, l'explication de l'évaluation et enfin, la présentation des résultats.

A3-2 Choix des images d'essai

Pour les essais en laboratoire, le contenu des images d'essai sera choisi selon la description faite au § 2.3 de la Partie 1.

Une fois le contenu choisi, les images d'essai sont préparées de manière à refléter les différentes configurations à l'étude ou la/les gamme(s) d'un ou de plusieurs paramètre(s). Lorsqu'on veut évaluer deux paramètres ou plus, les images peuvent être préparées de deux façons. Dans la première variante, chaque image représente un niveau d'un seul paramètre. Dans la seconde, chaque image représente un niveau de tous les paramètres examinés, mais image après image, chaque niveau de chaque paramètre apparaît avec chaque niveau de tous les autres paramètres. Les deux méthodes permettent de connaître précisément les résultats pour chaque paramètre. La dernière méthode permet également de déceler les interactions entre les différents paramètres (c'est-à-dire les effets non additifs).

A3-3 Séance de test

La séance de test comporte une série de présentations qui seront présentées dans un ordre aléatoire et, de préférence, dans un ordre différent pour chaque observateur. Lorsqu'on utilise un ordre de séquences aléatoire, il y a deux types de présentation: I (stimulus unique), et II (stimulus unique à répétitions multiples) qui sont présentés ci-dessous:

a) Les images ou séquences de test ne sont présentées qu'une seule fois pendant la séance de test; au début des premières séances, on introduira certaines séquences fictives (conformément à la description du § 2.7 de la Partie 1); l'expérimentateur veille habituellement à ce que la même image ne soit pas présentée deux fois de suite avec le même niveau de dégradation.

   Une présentation type comprend trois visualisations: une image d'adaptation gris moyen, une image de stimulus et de nouveau une image gris moyen. La durée de ces visualisations varie selon la tâche de l'observateur, le matériel (images fixes/images animées) et les options ou les paramètres considérés; mais 3, 10 et 10 s respectivement sont des durées courantes pour ces visualisations. L'avis ou les avis de l'observateur peuvent être recueillis pendant la visualisation de l'image de stimulus ou de la seconde image gris moyen.

b) Les images ou séquences de test sont présentées trois fois ce qui divise la séance de test en trois présentations, chacune d'elles ne comportant qu'une seule fois toutes les images ou séquences à tester; le début de chaque présentation est annoncé par l'apparition d'un message sur l'écran de contrôle (Présentation 1, par exemple); la première présentation sert à fixer l'opinion de l'observateur; les notes issues de cette présentation ne doivent pas être prises en considération dans les résultats du test; on obtient les notes attribuées aux images ou séquences en faisant la moyenne des notes attribuées pendant les deuxième et troisième présentations; l'expérimentateur veille habituellement à ce que les contraintes suivantes soient respectées concernant l'ordre aléatoire des images ou séquences à l'intérieur de chaque présentation:

   – une image ou séquence donnée n'occupe pas la même position dans les autres présentations;

   – une image ou séquence donnée n'est pas située immédiatement après la même image ou séquence dans les autres présentations.

Une présentation type comporte deux visualisations: une image de stimulus et une image gris moyen. La durée de ces visualisations varie selon la tâche de l'observateur, les séquences d'image de test et les opinions ou les facteurs considérés, mais les temps suggérés sont respectivement de 10 et 5 s. L'avis ou les avis de l'observateur ne peuvent être recueillis que pendant la visualisation de l'image gris moyen.

La variante II (stimulus unique à répétitions multiples) allonge manifestement la durée d'exécution d'une séance de test (45 s au lieu de 23 s pour chaque image ou séquence soumise au test); cela étant, les résultats de la variante I sont moins dépendants de l'ordre des images ou séquences au cours d'une séance.

Par ailleurs, des résultats expérimentaux montrent que la variante II autorise une fourchette d'environ 20% dans l'étalement des notations.

A3-4 Types de méthodes à un seul stimulus

Trois types de méthodes à un seul stimulus ont été généralement utilisés pour évaluer les systèmes de télévision.

A3-4.1 Méthodes utilisant une échelle d'évaluation par catégorie au moyen d'adjectifs

Dans ce cas, les observateurs attribuent à une image ou une séquence d'images une catégorie choisie parmi un ensemble de catégories définies d'un point de vue sémantique. Les catégories peuvent traduire la présence ou l'absence d'un attribut, par exemple, pour établir le seuil de dégradation. Les échelles par catégories permettant d'évaluer la qualité de l'image et la dégradation de l'image, ont été utilisées dans la plupart des cas; les échelles de l'UIT-R sont indiquées au Tableau 2-1. Dans le cas de l'affichage pendant l'exploitation, on utilise parfois des demi-notes. Des échelles permettant d'évaluer la lisibilité du texte, l'effort de lecture et l'utilité de l'image ont été utilisées dans des cas particuliers.

TABLEAU 2-1

Échelles de qualité et de dégradation de l'UIT-R

Échelle à cinq notes	
Qualité	**Dégradation**
5 Excellent	5 Imperceptible
4 Bon	4 Perceptible mais non gênant
3 Assez bon	3 Légèrement gênant
2 Médiocre	2 Gênant
1 Mauvais	1 Très gênant

Cette méthode aboutit, pour chaque condition, à une distribution des évaluations selon les catégories de l'échelle. La façon dont les réponses sont analysées dépend du jugement (détection, etc.) et de l'information recherchée (seuil de détection, rangs ou tendance moyenne des conditions, «distances» psychologiques entre les différentes conditions). Un grand nombre de méthodes d'analyse sont disponibles.

A3-4.2 Méthodes utilisant une échelle catégorielle numérique

Une méthode à un seul stimulus avec une échelle catégorielle numérique à 11 notes a été étudiée et comparée aux échelles graphiques et de rapports. Cette étude que décrit le Rapport UIT-R BT.1082 révèle, sur le plan de la sensibilité et de la stabilité, une nette préférence en faveur de cette méthode lorsque aucune référence n'est disponible.

A3-4.3 Méthodes n'utilisant pas une échelle d'évaluation par catégorie

Dans ce cas, les observateurs attribuent une valeur à chaque image ou séquence d'images présentée. Cette méthode a deux variantes.

Dans le cas d'une échelle continue, qui constitue une variante de la méthode par catégorie, l'observateur attribue à chaque image ou chaque séquence d'images un point situé sur une ligne tracée entre deux qualificatifs sémantiques (par exemple, les extrémités d'une échelle par catégorie comme au Tableau 2-1). Pour référence, l'échelle peut comporter d'autres qualificatifs, situés en des points intermédiaires. La distance jusqu'à l'une des extrémités de l'échelle sert d'indice pour chaque condition.

Dans le cas d'une échelle discrète, l'observateur attribue à chaque image ou séquence d'images une note qui reflète, pour un paramètre spécifique, le niveau de la qualité de l'image tel qu'il l'a apprécié (par exemple, la netteté de l'image). La gamme de notes utilisées peut être restreinte (par exemple, 0-100) ou non. Parfois, la note attribuée reflète le niveau apprécié en termes «absolus» (sans référence directe au niveau de qualité d'une quelconque autre image ou séquence d'images comme dans certaines formes de la méthode d'estimation des grandeurs). Dans d'autres cas, la note traduit le niveau apprécié par rapport au niveau considéré précédemment comme «type» (par exemple, méthode d'estimation des grandeurs, fractionnement et estimation par la méthode utilisant une échelle de rapport).

Dans un cas comme dans l'autre, on aboutit à une distribution des notes pour chaque condition d'essai. La méthode d'analyse utilisée dépend du type de jugement et de l'information requise (par exemple, rangs, tendance centrale, «distances» psychologiques).

A3-4.4 Mesures de la performance

Certains aspects des conditions normales d'observation peuvent être évalués en termes de «performance» des tâches purement externes (informations ciblées, lecture d'un texte, identification d'objets, etc.). Ainsi, une mesure de la performance portant par exemple sur la précision ou la rapidité avec laquelle ces tâches sont exécutées peut servir d'indice de l'image ou de la séquence d'images.

Les mesures de la performance conduisent à une distribution des notes appréciant la précision ou la rapidité pour chaque condition. L'analyse s'attache avant tout à établir les relations entre les conditions dans la tendance centrale (et dispersion) des notes et utilise souvent l'analyse de variance ou une technique analogue.

Annexe 4
de la Partie 2

Méthodes de comparaison de stimulus

Dans ce type de méthodes, on présente deux images ou séquences d'images à l'observateur qui fournit un indice de la relation entre les deux présentations.

A4-1 Mode opératoire général

Désigne la façon de définir et de choisir, conformément au § 2 de la Partie 1, les conditions d'observation, les signaux source, la gamme de conditions et l'ancrage, les observateurs, l'explication de l'évaluation et enfin, la présentation des résultats.

A4-2 Choix du matériel d'essai

Les images ou séquences d'images utilisées sont produites de la même façon que dans les méthodes à un seul stimulus. Les images ou séquences d'images ainsi obtenues sont ensuite combinées pour former les paires utilisées dans les essais d'évaluation.

A4-3 Séance d'évaluation

L'évaluation fera intervenir soit un seul écran d'évaluation soit deux bien synchronisés et se déroulera généralement comme dans le cas des méthodes à un seul stimulus. Si on utilise un seul écran d'évaluation, la présentation élémentaire comportera un stimulus supplémentaire identique en durée au premier. Dans ce cas, on fera bien de s'assurer au fil des essais, que les deux membres d'une paire apparaissent un même nombre de fois en première et en seconde position. Si on utilise deux écrans d'évaluation, les images de stimulus sont présentées simultanément.

Les méthodes de comparaison de stimulus permettent d'évaluer plus complètement les relations existant entre les conditions lorsque les évaluations portent sur toutes les paires possibles de conditions. Toutefois, s'il faut un trop grand nombre d'observations, on peut répartir les observations entre les observateurs, ou utiliser un échantillon de toutes les paires possibles.

A4-4 Types de méthodes de comparaison de stimulus

Trois types de méthodes de comparaison de stimulus ont été utilisés en vue d'évaluer des systèmes de télévision.

A4-4.1 Méthodes utilisant une échelle d'évaluation par catégorie au moyen d'adjectifs

Dans ce genre de méthode, les observateurs estiment la relation entre les membres d'une paire en attribuant une catégorie choisie parmi un ensemble de catégories définies d'un point de vue sémantique. Ces catégories peuvent indiquer la présence de différences perceptibles (par exemple, IDENTIQUE, DIFFÉRENT), la présence et le degré de différences perceptibles (par exemple, MOINS, IDENTIQUE, PLUS) ou des appréciations de l'importance et du degré des différences. L'échelle comparative de l'UIT-R est indiquée au Tableau 2-2.

TABLEAU 2-2

Échelle de comparaison

–3	Beaucoup moins bon
–2	Moins bon
–1	Légèrement moins bon
0	Identique
+1	Légèrement mieux
+2	Mieux
+3	Beaucoup mieux

Cette méthode conduit, pour chaque paire de conditions, à une distribution des évaluations subjectives sur les catégories de l'échelle. La façon dont les réponses sont analysées dépend de l'appréciation (par exemple, différence) et de l'information requise (par exemple, différences juste perceptibles, classement des conditions, «distances» entre les conditions, etc.).

A4-4.2 Méthodes n'utilisant pas une échelle d'évaluation par catégorie

Dans ce genre de méthode, les observateurs attribuent une valeur à la relation entre les membres d'une paire d'évaluations subjectives. Cette méthode présente deux variantes:

– Dans le cas d'une échelle continue, l'observateur attribue à chaque relation un point situé sur une ligne tracée entre deux qualificatifs (par exemple, IDENTIQUE-DIFFÉRENT ou les extrémités d'une échelle par catégorie comme dans le Tableau 2-2). Les échelles peuvent comporter d'autres qualificatifs de référence situés en des points intermédiaires. La distance qui sépare le point de l'extrémité de la ligne sert de référence pour chaque paire de conditions.

– Dans la seconde variante, l'observateur attribue à chaque relation une note qui reflète le niveau de l'image tel qu'il l'a perçu, cela pour un paramètre précis (par exemple, la différence de qualité). La gamme des notes utilisées peut être limitée ou non. La note attribuée peut décrire la relation en termes «absolus» ou en termes d'une paire «type».

Dans les deux cas, on obtient une distribution des valeurs pour chaque paire de conditions. La méthode d'analyse dépend de la nature de l'appréciation portée et de l'information requise.

A4-4.3 Mesures de la performance

Dans certains cas, les mesures de la performance peuvent être obtenues à partir de méthodes de comparaison de stimulus. Dans la méthode du choix forcé, chaque paire d'images est préparée de telle sorte que l'une des images présente un niveau spécifique d'un attribut (par exemple, dégradation) alors que l'autre présente un niveau différent de ce même attribut ou ne présente pas cet attribut. L'observateur est prié d'indiquer l'image qui présente le niveau le plus élevé/le moins élevé de l'attribut ou l'image qui ne présente pas l'attribut; la précision et la rapidité de la performance servent à mesurer la relation entre les membres de la paire.

Annexe 5
de la Partie 2

Évaluation continue de la qualité avec stimulus unique (SSCQE)

La compression du signal de télévision numérique va entraîner des dégradations pour la qualité de l'image, dégradations qui dépendent de la scène et varient en fonction du temps. Même sur de courtes séquences d'enregistrements vidéo numériques, la qualité peut varier dans des proportions importantes selon le contenu de la scène, et les dégradations peuvent être très brèves. Les méthodologies classiques de l'UIT-R ne permettent pas à elles seules d'évaluer ce type de séquence d'image de test. Par ailleurs, la méthode à double stimulus utilisées pour les tests en laboratoire ne reproduit pas les conditions du téléspectateur à son domicile qui, lui, ne dispose que d'un seul stimulus. On a donc jugé utile de mesurer la qualité subjective de séquences vidéo numériques de façon continue, les sujets visualisant les séquences d'image test une seule fois, sans référence source.

La technique d'évaluation continue de la qualité avec stimulus unique (SSCQE) a donc a été mise au point et testée.

A5-1 Dispositif d'enregistrement et configuration

Un système d'enregistrement électronique connecté à un ordinateur sera utilisé pour enregistrer l'évaluation continue de la qualité faite par les sujets. Ce dispositif aura les caractéristiques suivantes:

– mécanisme à glissière sans position de rappel;
– course: 10 cm;
– fixe ou pouvant être installé sur un bureau;
– échantillons enregistrés deux fois par seconde.

A5-2 Forme générale du protocole de test

On présentera aux sujets des séances de test du format suivant:

– *segment de programme (SP)*: correspond à un type de programme (sport, journal télévisé, dramatique, par exemple) traité conformément à l'un des paramètres de qualité (PQ) à évaluer (par exemple, débit binaire); chaque segment durera au moins 5 min;
– *séance de test (ST)*: série d'une ou de plusieurs combinaisons différentes SP/PQ non séparées et ordonnées de façon pseudo-aléatoire. Chaque séance de test contient au moins une fois tous les segments de programme et paramètres de qualité, mais pas nécessairement toutes les combinaisons SP/PQ; chaque ST durera entre 30 et 60 minutes;
– *présentation de test (PT)*: correspond à l'intégralité d'un test. Elle peut être divisée en ST pour respecter les impératifs en ce qui concerne la durée maximale du test et pour évaluer la qualité de toutes les paires SP/PQ. Si le nombre de ces paires est limité, une présentation peut comporter plusieurs fois la même ST afin que le test dure assez longtemps.

Si l'on veut évaluer la qualité de service, on peut introduire une séquence audio. Dans ce cas, on apportera le même soin au choix de la séquence audio d'accompagnement et de la séquence vidéo avant de procéder au test.

Le format de test le plus simple se composera d'un seul SP et d'un seul PQ.

A5-3 Paramètres d'observation

Les conditions d'observation seront celles qui sont actuellement indiquées dans la Partie 1 ou les conditions prévues pour les applications décrites dans la Partie 3.

A5-4 Échelles d'évaluation

Les sujets sont avertis dans les instructions du test que la course du mécanisme à glissière correspond à l'échelle de qualité continue décrite au § A1-4 de la Partie 2.

A5-5 Observateurs

On utilisera au moins quinze sujets, non spécialistes, conformément aux conditions actuellement recommandées au § 2.5 de la Partie 1.

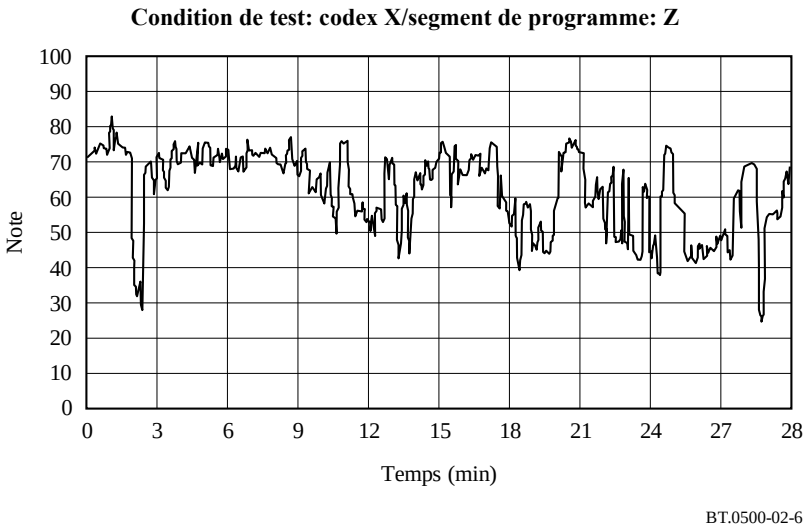
A5-6 Directives à l'intention des observateurs

Dans le cas d'une évaluation de la qualité de service (avec séquence audio d'accompagnement), il sera demandé aux observateurs de juger la qualité d'ensemble et non la qualité vidéo seulement.

A5-7 Présentation de données, traitement et présentation des résultats

Les données de toutes les séances de test seront regroupées et exploitées. On peut donc tracer une courbe unique représentant des notes moyennes de qualité en fonction du temps $q(t)$; ce sera la moyenne de toutes les notes de qualité données par les observateurs par segment de programme, paramètre de qualité ou par séance de test entière (voir l'exemple illustré sur la Fig. 2-6).

FIGURE 2-6



Toutefois, les différences de temps de réaction entre téléspectateurs peuvent influencer les résultats de l'évaluation si seule la moyenne sur un segment de programme est calculée. Des études destinées à évaluer l'incidence du temps de réaction sur la note de qualité obtenue sont actuellement en cours.

Ces données peuvent être converties en un histogramme de probabilité, $P(q)$, d'occurrence du niveau de qualité q (voir l'exemple illustré à la Fig. 2-7).

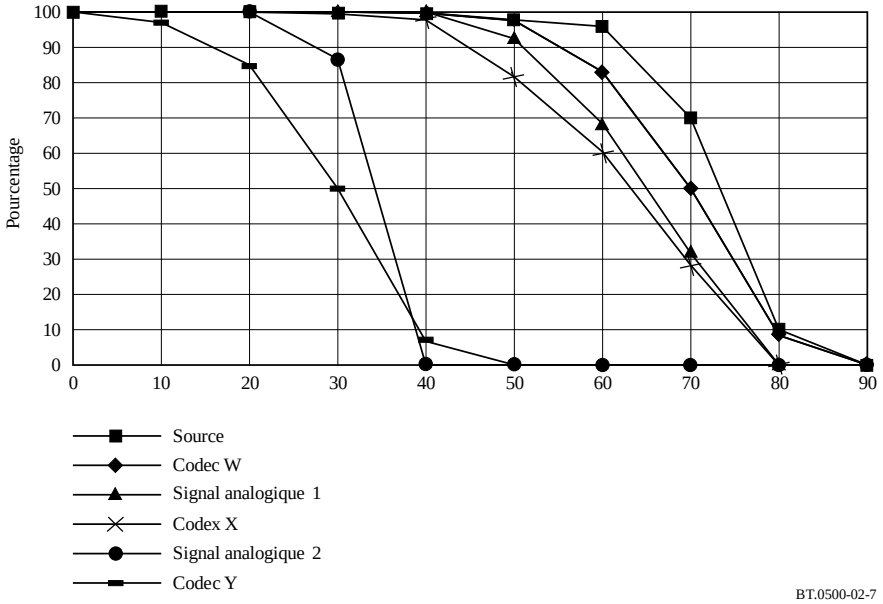
A5-8 Étalonnage des résultats d'évaluation continue de la qualité et obtention d'une évaluation d'ensemble de la qualité unique

On a pu constater que des erreurs systématiques imputables à la mémoire des sujets peuvent apparaître lors de longues séances DSCQS d'évaluation globale de la qualité de séquences d'enregistrements vidéo numériques mais on a vérifié récemment que ces erreurs ne sont pas significatives dans des évaluations DSCQS d'extraits d'enregistrements vidéo de 10 s. Une deuxième étape possible du processus d'évaluation continue de la qualité à stimulus unique (SSCQE), actuellement à l'étude, consisterait donc à étalonner l'histogramme de qualité, à l'aide de la méthode DSCQS existante, sur des échantillons de 10 s représentatifs extraits des données de l'histogramme.

Les méthodologies classiques que l'UIT-R a utilisées dans le passé ont permis d'obtenir des notations globales de la qualité de séquences de télévision. Des expériences ont été faites pour examiner la relation existant entre l'évaluation continue de la qualité d'une séquence vidéo codée et une évaluation globale de la qualité du même segment. On a déjà constaté que la mémoire humaine peut être trompeuse et fausser les notations de la qualité si des dégradations perceptibles apparaissent approximativement dans les 10 à 15 dernières secondes de la séquence, mais on a également constaté que ces effets trompeurs de la mémoire humaine pouvaient être modélisés sous forme d'une fonction exponentielle décroissante. Une troisième étape possible de la méthodologie SSCQE consisterait donc à traiter les résultats de ces évaluations continues de la qualité pour obtenir une mesure globale de la qualité correspondante. Cela est actuellement à l'étude.

FIGURE 2-7

Moyenne des notes données pendant les séquences de notation concernant le segment de programme Z



Annexe 6
de la Partie 2

Méthode d'évaluation continue à double stimulus simultané (SDSCE)

L'UIT-R a conçu une méthode d'évaluation continue parce que les méthodes précédentes ne convenaient pas parfaitement à la mesure de la qualité vidéo des systèmes de compression numérique. En effet, ces méthodes présentaient des inconvénients majeurs liés à l'occurrence de perturbations associées au contexte dans les images numériques affichées. Dans les autres protocoles, la durée d'observation des séquences vidéo soumises à l'évaluation est en général limitée à 10 s, ce qui de toute évidence n'est pas suffisant pour que l'observateur porte un jugement représentatif de celui qui aurait le spectateur, dans le service réel. Les perturbations numériques sont étroitement liées au contenu spatial et temporel de l'image source. Cela est vrai non seulement pour les systèmes de compression mais aussi pour la protection contre les erreurs de systèmes de transmission numérique. Avec les méthodologies normalisées précédentes, il était très difficile de choisir des séances vidéo représentatives, ou du moins d'évaluer leur représentativité. C'est la raison pour laquelle l'UIT-R a introduit la méthode SSCQE qui permet de mesurer la qualité vidéo sur des séquences plus longues, représentative du contenu vidéo et des statistiques d'erreur. Afin de reproduire des conditions d'observation aussi proches que possible de situations réelles, on n'utilise pas de référence dans la méthode SSCQE.

Lorsqu'il faut évaluer la fidélité, il faut introduire des conditions de référence. La méthode SDSCE a été élaborée à partir de la méthode SSCQE, avec de légères modifications en ce qui concerne la manière de présenter des images aux sujets et l'échelle de notation. Cette méthode a été proposée au Groupe MPEG pour évaluer l'invulnérabilité aux erreurs à un débit binaire très faible, mais elle peut être employée avec de bons résultats dans tous les cas où la fidélité d'informations visuelles affectées par une dégradation variable dans le temps doit être évaluée.

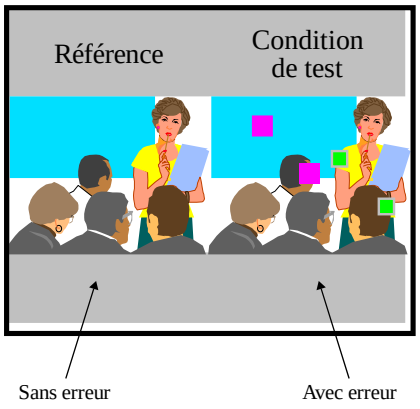
La nouvelle technique SDSCE décrite ci-dessous a donc été mise au point et testée.

A6-1 Procédure de test

Le groupe de sujets observe deux séquences en même temps: l'une est la séquence de référence, l'autre correspond aux conditions de test. Si les deux séquences sont présentées en format d'image standard (SIF, *standard image format*) ou en un format plus petit, elles peuvent être affichées en parallèle sur le même écran, autrement il faut utiliser deux écrans placés côte à côte (voir la Fig. 2-8).

FIGURE 2-8

Exemple de format d'affichage



On demande aux sujets de constater les différences entre les deux séquences et de juger la fidélité des informations vidéo en déplaçant la glissière du mécanisme de notation. Lorsque la fidélité est parfaite, la glissière devrait se trouver au sommet de l'échelle de notation (soit 100), et lorsque la fidélité est nulle, elle devrait se trouver au bas de l'échelle (soit 0).

Les sujets savent quelle est la séquence de référence et ils doivent faire connaître leur opinion tout en observant les séquences, pendant toute la durée de celles-ci.

A6-2 Les différentes phases

La *phase de préparation* est essentielle dans cette méthode de test pour que les sujets comprennent bien ce qu'ils doivent faire. Il faut leur donner des instructions écrites pour être sûr que tous reçoivent exactement les mêmes informations. Ces instructions doivent comprendre des explications sur ce qu'ils vont voir, ce qu'ils doivent évaluer (c'est-à-dire la différence de qualité) et sur les moyens à utiliser pour exprimer leur opinion. Il faut répondre aux questions que pourraient poser les sujets afin d'éviter autant que possible toute opinion partiale induite par le responsable du test.

Après avoir communiqué les instructions, il convient d'organiser une *séance de démonstration* afin que les sujets puissent se familiariser tant avec les procédures de notation qu'avec les types de dégradation.

Enfin, il convient de procéder à une simulation de test comprenant un certain nombre de conditions représentatives. Les séquences devraient être différentes de celles employées dans le test et passer l'une après l'autre, sans interruption.

Lorsque la *simulation de test* est terminée, l'expérimentateur doit vérifier, surtout lorsque les séquences de conditions de test sont identiques à celles de référence, que les évaluations sont proches de cent (c'est-à-dire qu'aucune différence n'a été perçue); si les sujets déclarent avoir perçu des différences, l'expérimentateur doit alors reprendre les explications et la simulation de test.

A6-3 Caractéristiques du protocole de test

Les définitions ci-après sont utilisées pour décrire le protocole de test:

– *Segment vidéo (SV)*: correspond à une séquence vidéo.

– *Condition de test (CT)*: peut être soit un processus vidéo spécifique, une condition de transmission ou les deux. Chaque SV doit être traité conformément à une CT au moins. En outre, les références doivent être ajoutées à la liste des conditions de test, afin de créer des paires référence/référence à évaluer.

– *Séance (S)*: série de paires différentes SV/CT non séparées et ordonnées de façon pseudo-aléatoire. Chaque séance contient au moins une fois tous les SV et les CT, mais pas nécessairement toutes les combinaisons SV/CT.

– *Présentation de test (PT)*: série de séances comprenant toutes les combinaisons SV/CT. Toutes les combinaisons SV/CT doivent être notées par le même nombre d'observateurs (mais pas nécessairement par les mêmes observateurs).

– *Période de notation*: chaque observateur est prié de noter de manière continue pendant une séance.

– *Segment de notation:* segment de 10 s de notation; tous les segments de notation sont obtenus par l'utilisation de groupes de 20 notes consécutives (l'équivalent de 10 s) sans aucun chevauchement.

A6-4 Traitement des données

Une fois que le test a été mené à bien, on établit un (ou plusieurs) fichier(s) de données regroupant toutes les notes des différentes séances (S) qui représentent l'ensemble du matériel de notation de la présentation de test (PT). On peut effectuer un premier contrôle de la validité des données en vérifiant que chaque paire SV/CT a été traitée et qu'un nombre équivalent de notes a été attribué à chacune des paires.

Les données, collectées durant l'exécution des tests effectués conformément à ce protocole, peuvent être traitées de trois manières différentes:

– analyse statistique de chaque SV distinct;

– analyse statistique de chaque CT distincte;

– analyse statistique globale de toutes les paires SV/CT.

Pour chaque cas, il faut procéder à une analyse en plusieurs étapes:

– Les moyennes et les écarts types sont calculés pour chaque notation par cumul des observations.

– La moyenne et l'écart type sont calculés pour chaque segment de notation, comme illustré à la Fig. 2-9. Les résultats de cette étape peuvent être représentés par un diagramme temporel (Fig. 2-10).

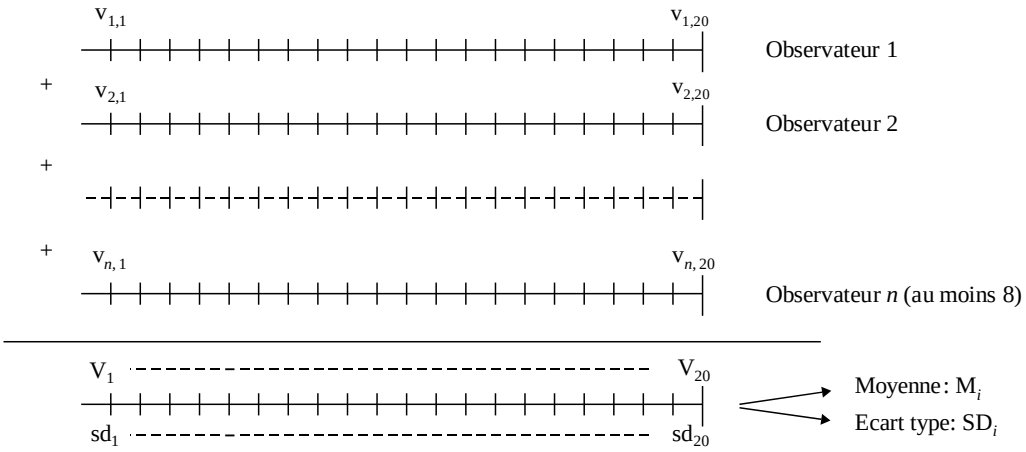
– On analyse la répartition statistique des moyennes calculées à l'étape précédente (c'est-à-dire celles correspondant à chaque segment de notation) et leur fréquence d'occurrence. Afin d'éviter un effet de rémanence dû à la combinaison SV × CT précédente, on ne tient pas compte des 10 premières secondes de notation pour chaque échantillon SV × CT.

– On calcule les caractéristiques globales de gêne par cumul des fréquences d'occurrence. Il faut tenir compte dans ce calcul des intervalles de confiance, comme indiqué à la Fig. 2-11. Des caractéristiques globales de gêne correspondent à cette fonction de répartition statistique cumulative en indiquant la relation entre les moyennes pour chaque segment de notation et leur fréquence cumulative d'occurrence.

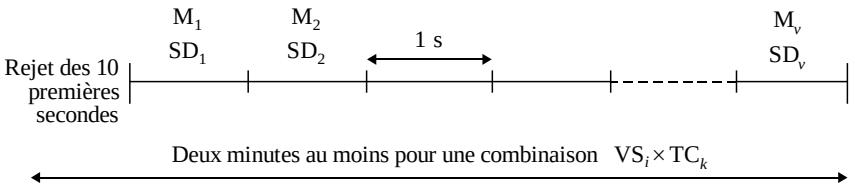
FIGURE 2-9

Traitement des données

a) Calcul de note moyenne, V et de l 'écart type, SD, par instant de vote pour les observateurs pour chaque séquence de notation de chaque combinaison VS × TC



b) Calcul de la moyenne, M et de l 'écart, SD, par séquence de notation d 'une seconde pour chaque combinaison VS × TC



BT.0500-02-9

A6-5 Fiabilité des sujets

On peut évaluer qualitativement la fiabilité des sujets en analysant leur comportement lorsqu'on leur montre des paires référence/référence. Dans ce cas, les sujets sont censés donner des évaluations très proches de 100. On peut ainsi constater qu'ils ont au moins compris ce que l'on attendait d'eux et qu'ils n'ont pas donné de note de manière aléatoire.

En outre, on peut contrôler la fiabilité des sujets au moyen de procédures similaires à celle décrite au § A1-2.3.2 de l'Annexe 1 de la Partie 1 de la méthode SSCQE.

Dans le cadre de la procédure SDSCE, la fiabilité des notes dépend des deux paramètres suivants:

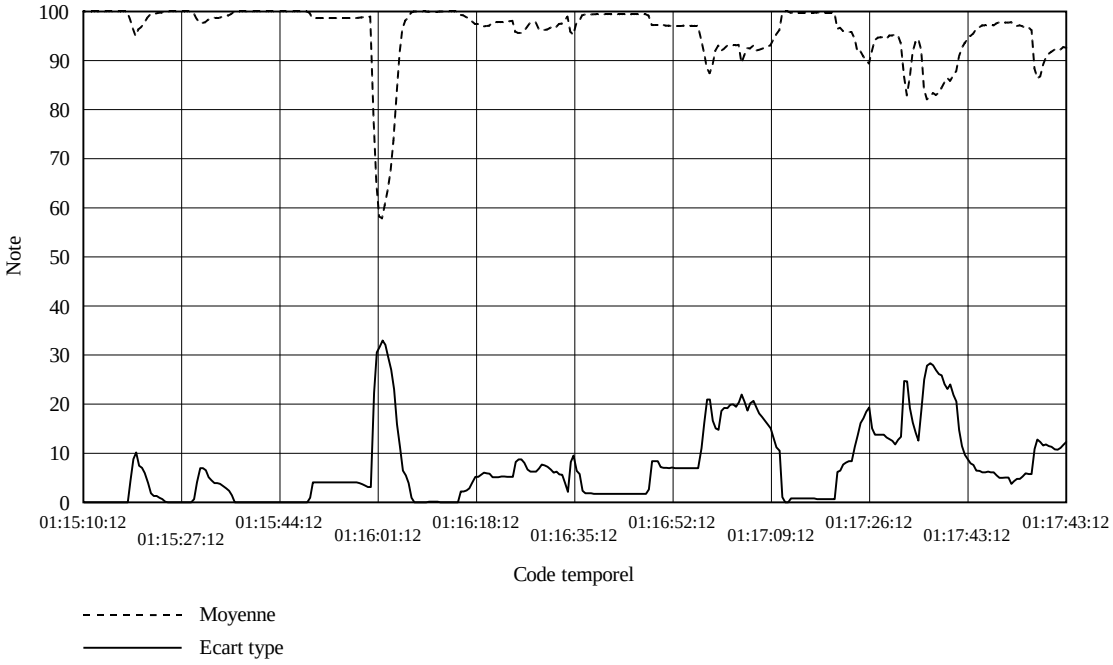
Décalages systématiques: pendant un test, un observateur peut être trop généreux ou trop prudent et même ne pas avoir compris les procédures de notation (signification de l'échelle de notation par exemple). Cela peut donner une série de notes systématiquement plus ou moins décalées, voire extrêmes, par rapport aux séries moyennes.

Inversions locales: comme dans d'autres procédures de test courantes, les observateurs peuvent quelquefois noter sans observer ou suivre attentivement la qualité des séquences affichées. Dans ce cas, la courbe totale des notes peut se trouver relativement dans la fourchette moyenne, mais on peut néanmoins constater des inversions locales.

Ces deux effets indésirables (comportement atypique et inversions) doivent être évités. Il est de toute évidence très important de former les participants, mais on devrait pouvoir utiliser un moyen permettant de détecter, et, si nécessaire, d'écarter les observateurs incohérents. La présente Recommandation décrit un processus en deux étapes qui permet de filtrer les participants.

FIGURE 2-10

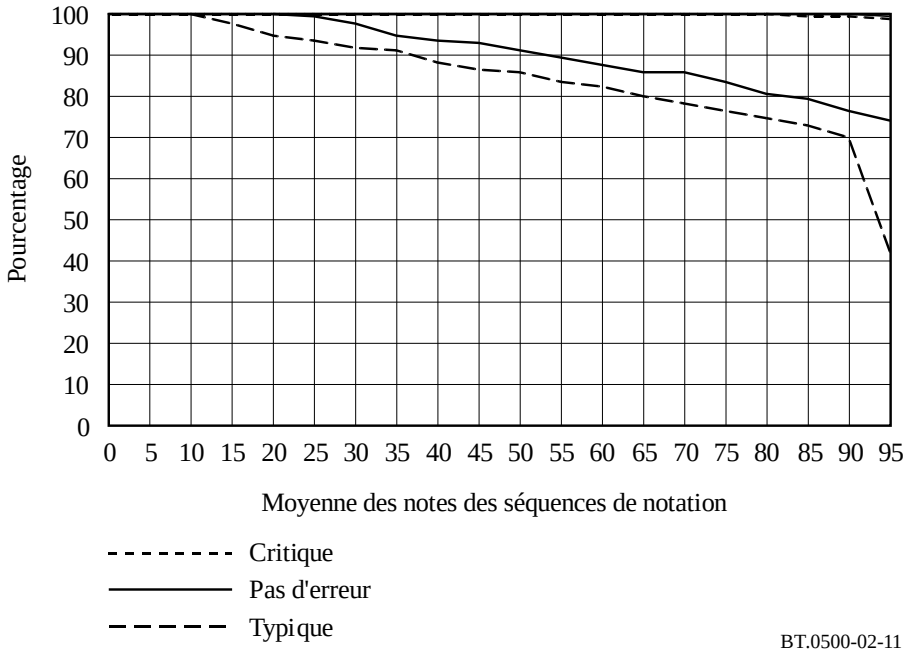
Diagramme temporel brut



BT.0500-2-10

FIGURE 2-11

Caractéristiques globales de gêne calculées à partir des répartitions statistiques et comprenant l'intervalle de confiance



BT.0500-02-11

Annexe 7
de la Partie 2

Évaluation subjective de la qualité vidéo multimédia (SAMVIQ)

A7-1 Introduction

La méthode SAMVIQ d'évaluation de la qualité utilise une échelle de qualité continue pour l'évaluation de la qualité intrinsèque des séquences vidéo. Chaque observateur déplace un curseur sur une échelle continue graduée de 0 à 100 avec 5 plages de qualité (excellente, bonne, assez bonne, médiocre, mauvaise).

Dans la méthode SAMVIQ, l'observateur a accès à plusieurs versions d'une même séquence. Lorsqu'il a évalué toutes les versions, il peut ensuite accéder au contenu de la séquence suivante.

L'observateur peut choisir aléatoirement chacune des différentes versions par le biais d'une interface graphique informatique. Il a la possibilité d'arrêter chaque version d'une séquence, de la revoir et d'en modifier la note. Cette méthode comporte une séquence de référence (c'est-à-dire non traitée) explicite ainsi que plusieurs versions de la même séquence qui incluent à la fois des séquences traitées et une séquence non traitée (c'est-à-dire une référence cachée). Chaque version d'une séquence est affichée isolément et évaluée sur une échelle de qualité continue analogue à celle utilisée dans la méthode DSCQS. Du point de vue fonctionnel, la méthode ressemble donc beaucoup à une méthode à un seul stimulus avec accès aléatoire, mais un observateur peut visualiser la référence explicite chaque fois qu'il le souhaite, ce qui rend cette méthode analogue à une méthode utilisant une référence.

La méthode SAMVIQ utilise une échelle de qualité continue pour l'évaluation de la qualité intrinsèque des séquences vidéo. Chaque observateur déplace un curseur sur une échelle continue graduée de 0 à 100 avec cinq plages de qualité (excellente, bonne, assez bonne, médiocre, mauvaise).

L'évaluation de la qualité est réalisée *scène par scène* (voir la Fig. 2-12), avec *une référence explicite, une référence cachée* et *divers algorithmes*.

Pour rendre la méthode plus facilement compréhensible, on définit les termes spécifiques qui suivent:

Scène: contenu audiovisuel

Séquence: scène avec traitement combiné ou sans traitement

Algorithme: une ou plusieurs techniques de traitement d'image.

A7-2 Référence explicite, référence cachée et algorithmes

Une méthode d'évaluation comporte généralement des ancres de qualité afin de stabiliser les résultats. Deux ancres de qualité élevée sont utilisées dans la méthode SAMVIQ pour les raisons précisées ci-après. La réalisation de plusieurs essais a montré que l'écart type des notes est minimalisé si on utilise une *référence explicite* plutôt qu'une référence cachée ou qu'aucune référence. En particulier, pour évaluer les performances d'un codec, il vaut mieux utiliser une référence explicite pour obtenir des résultats les plus fiables possibles. Une *référence cachée* est également utilisée afin d'évaluer la qualité intrinsèque de la référence; on n'utilise pas la référence explicite car la présentation est anonyme de même que les séquences traitées. La désignation explicite de «référence» a une influence sur environ 30% des observateurs, qui donnent la note la plus élevée (100) à la référence explicite. Cette note est totalement différente de la note correspondante donnée pour la référence cachée. Il est à noter que lorsque aucune référence n'est disponible, l'essai reste possible mais l'écart type est nettement plus grand.

La méthode SAMVIQ convient bien dans un contexte multimédia car il est possible de combiner différents éléments du traitement d'image (par exemple type de codec, format d'image, débit, mise à jour temporelle, zoomage, etc.). Un *algorithme* désigne l'un de ces éléments ou une combinaison de ces éléments.

A7-3 Conditions d'essai

La variation de la criticité pendant une scène est limitée car on choisit des contenus homogènes en suivant les mêmes règles que celles qui sont implicitement utilisées par les autres méthodes donnant une note globale (par exemple, les méthodes à un seul stimulus). Une durée maximale de visualisation de séquence de 10 ou 15 s est alors suffisante pour obtenir une note de qualité stabilisée et fiable. Il convient d'utiliser les décodeurs-lecteurs propriétaires, ou une copie d'écran de leur sortie, afin de maintenir la qualité d'affichage voulue.

A7-4 Organisation d'un essai

a) L'essai est réalisé scène par scène, comme décrit sur la Fig. 2-12.

b) Pour une scène donnée, il est possible de visualiser et de noter n'importe quelle séquence dans n'importe quel ordre. Chaque séquence peut être visualisée et notée plusieurs fois.

c) D'une scène à l'autre, l'accès aux séquences est randomisé, ce qui empêche les observateurs de tenter de voter de manière identique selon un ordre établi. En réalité, pour un essai donné, l'ordre des algorithmes reste le même afin de simplifier l'analyse et la présentation des résultats. Seul l'accès correspondant à partir d'un bouton identique est randomisé.

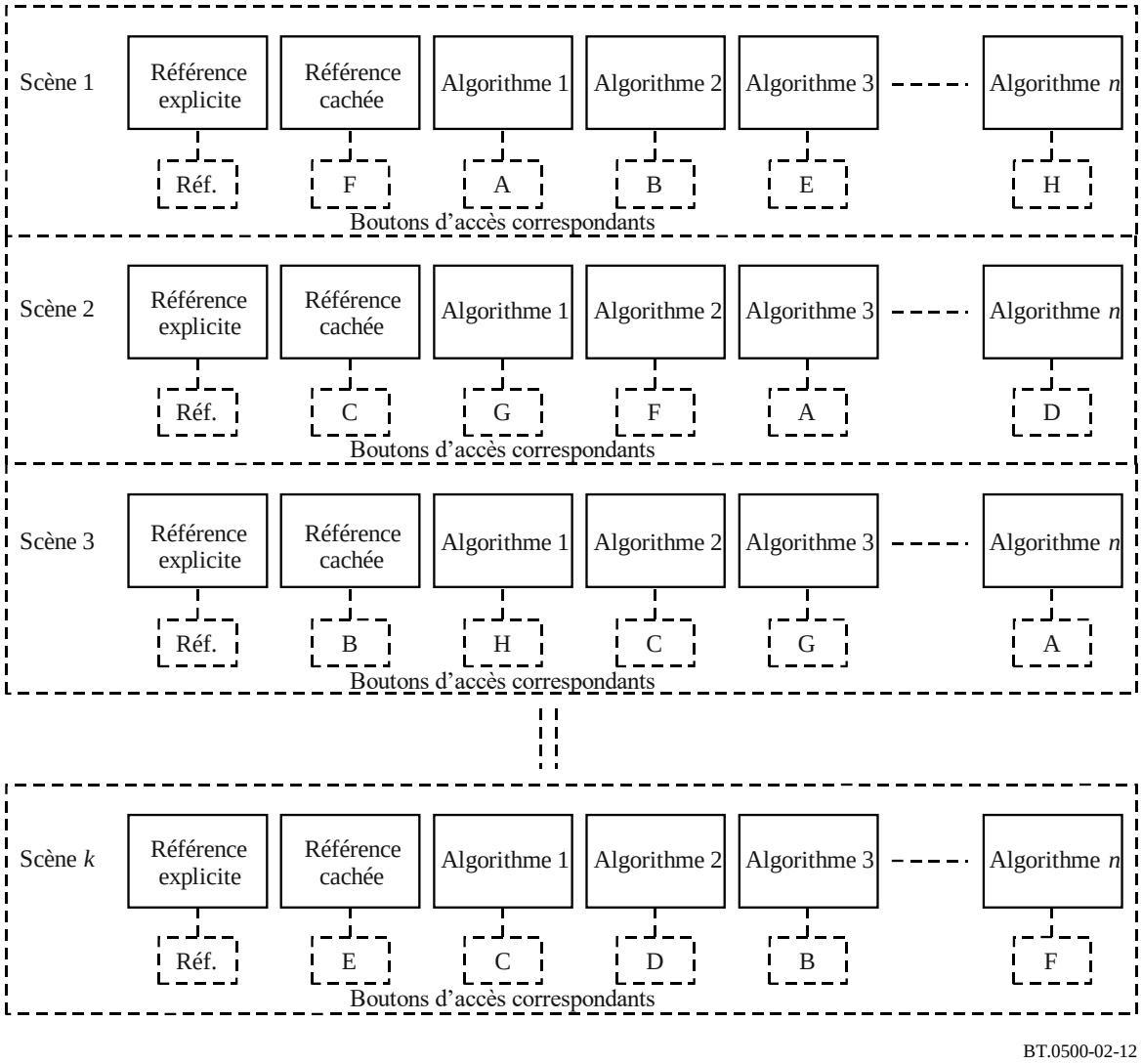
d) Lors de la première visualisation, la séquence considérée doit être visualisée en totalité avant d'être notée, afin d'éviter que la séquence soit notée et arrêtée immédiatement.

e) Pour pouvoir passer à la scène suivante, toutes les séquences de la scène considérée doivent avoir été notées.

f) L'essai est terminé quand toutes les séquences de toutes les scènes ont été notées.

FIGURE 2-12

Exemple d'organisation d'un essai pour la méthode SAMVIQ



BT.0500-02-12

La méthode SAMVIQ est mise en œuvre sous forme logicielle. En plus des boutons d'accès montrés sur la Fig. 2-12, des boutons «lecture», «arrêt», «scène suivante» et «scène précédente» sont nécessaires pour permettre à l'observateur de gérer la présentation des différentes scènes (voir par exemple le § A7-6). Lorsqu'une note a été donnée par l'observateur, elle devrait apparaître au-dessous du bouton d'accès correspondant à la scène considérée. Lorsque toutes les versions différentes d'une séquence ont été évaluées, l'observateur a toujours la possibilité de comparer les notes et, si nécessaire, de les modifier. Il n'est pas nécessaire de revoir la totalité de la séquence considérée car les grandes différences ont déjà été mises en avant au cours de la première visualisation.

A7-5 Présentation et analyse des données

A7-5.1 Informations relatives à l'essai

Des informations précises relatives à l'environnement de l'essai sont nécessaires pour pouvoir reproduire un essai ou comparer les résultats de différents essais. Il est donc suggéré d'indiquer les informations relatives à l'environnement d'essai décrites dans le Tableau 2-3.

TABLEAU 2-3

Informations relatives à l'essai

Nom de la méthode	
Technologie d'affichage	
Nom de référence de l'affichage	
Niveau de luminance de crête (cd/m²)	
Niveau de luminance du noir (cd/m²)	
Niveau du noir: PLUGE (seuil de perception de la distance entre le niveau du noir et le niveau du noir supra = 8). Sinon, indiquer la valeur seuil	
Niveau de luminance de l'arrière-plan (cd/m²)	
Éclairement (lux)	
Distance d'observation: – Pas de valeur imposée: devant l'affichage – Valeur imposée: nH	
Taille de l'affichage (diagonale en pouces)	
Rapport largeur/hauteur de l'affichage	
Format de l'affichage (nombre de colonnes et de lignes)	
Format d'entrée de l'image (nombre de colonnes et de lignes)	
Format de sortie de l'image(1) (nombre de colonnes et de lignes)	
Température du blanc: D_{65} sinon Coordonnées du blanc (x, y)	
Nombre d'observateurs effectifs	

(1) Cette information est nécessaire lorsque l'image d'entrée est traitée, par exemple si son échelle est changée, au moment de son affichage.

Les caractéristiques de l'affichage peuvent avoir une incidence sur les résultats de l'essai. D'autres informations (par exemple, la réponse en luminance (fidélité du gamma) et les couleurs primaires) devraient être nécessaires pour les affichages sur écran plat.

Les caractéristiques des séquences vidéo sont importantes pour concevoir un essai ou pour expliquer ses résultats. Il est suggéré d'indiquer les caractéristiques spatio-temporelles décrites à l'Annexe 1 de la Partie 1. Il convient de tenir compte de ces informations pour la collecte de séquences d'essai dans la bibliothèque de matériel vidéo à utiliser pour l'évaluation subjective de la qualité vidéo dans les applications multimédias.

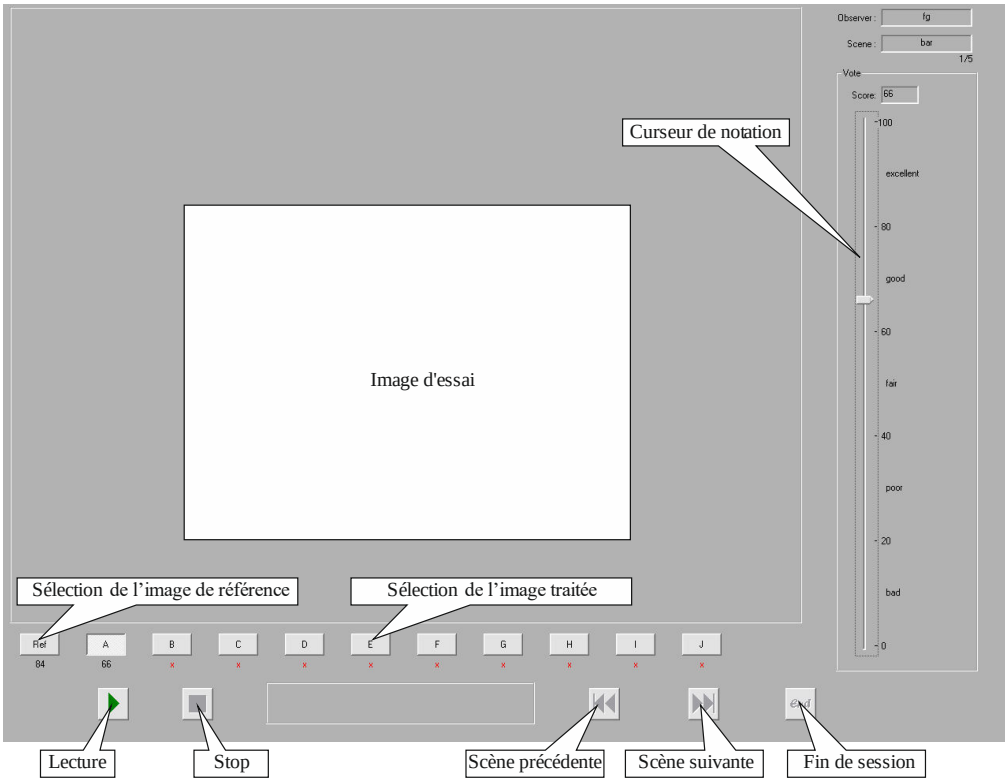
A7-5.2 Méthodes d'analyse

Les méthodes d'analyse sont celles décrites dans l'Annexe 1 de la Partie 1.

A7-5.3 Sélection des observateurs

Pour la méthode SAMVIQ, la procédure de selection est celle décrite au § A1-2.3.3 de l'Annexe 1 de la Partie 1.

A7-6 Exemple d'interface pour la méthode SAMVIQ (pour information)



Annexe 8
de la Partie 2

Protocole d'observation par des spécialistes (EVP) pour l'évaluation subjective de la qualité des séquences vidéo

La présente Annexe décrit la méthode d'évaluation subjective de la qualité vidéo d'images animées au moyen du protocole d'observation par des spécialistes (EVP, *expert viewing protocol*), avec la participation d'un nombre réduit d'observateurs, tous choisis parmi les spécialistes du domaine correspondant du traitement vidéo.

A8-1 Montage de laboratoire

A8-1.1 Choix de l'écran et montage

L'écran utilisé devrait être un écran plat ayant les caractéristiques types d'applications professionnelles (par exemple, studio de radiodiffusion ou car-régie); sa dimension en diagonale pourra être comprise entre 22 pouces (minimum) et 40 pouces (recommandée), mais elle pourra aller jusqu'à 50 pouces ou plus pour l'évaluation de systèmes d'images avec une résolution de TVHD ou plus élevée.

Il est possible d'utiliser une partie réduite de la zone active d'affichage de l'écran; dans ce cas, la zone située autour de la partie active de l'écran devrait être mise à «gris moyen». Dans cette configuration,

il ne devrait pas être accepté que la résolution de l'écran soit réglée sur une résolution autre que celle d'origine.

L'écran devrait permettre un montage et un étalonnage de la luminance et des couleurs appropriés au moyen d'un luxmètre professionnel. L'étalonnage de l'écran devrait être conforme aux paramètres définis dans la Recommandation applicable pour le test effectué.

A8-1.2 Distance d'observation

La distance d'observation devrait être choisie en fonction de la résolution de l'écran et de la hauteur de la partie active de l'écran, conformément à la distance d'observation nominale définie au § 2.1.3.2 de la Partie 1 ou être plus courte, selon les exigences concernant les mauvaises conditions d'observation.

A8-1.3 Conditions d'observation

Une expérience menée selon le protocole d'observation par des spécialistes (EVP) ne devrait pas nécessairement se dérouler dans un laboratoire de test, mais il est important que l'emplacement choisi soit protégé des perturbations sonores et/ou visuelles (on peut par exemple utiliser un bureau ou une salle de réunion au calme).

Il convient d'éliminer tout reflet sur l'écran provenant d'une source de lumière directe ou indirecte; l'éclairage ambiant devrait être faible, au niveau minimum suffisant pour remplir les feuilles de notation (le cas échéant).

Le nombre de spécialistes assis devant l'écran pourra varier en fonction de la taille de l'écran, l'objectif étant de garantir que le rendu d'image et l'exposition aux stimuli soient identiques pour tous les observateurs.

A8-2 Observateurs

Les observateurs participant à une expérience EVP devraient être des spécialistes du domaine à l'étude.

Les observateurs ne devraient pas nécessairement être sélectionnés pour leur acuité visuelle ou leur perception des couleurs, mais devraient être choisis parmi des personnes qualifiées.

Il devrait y avoir au minimum neuf observateurs différents.

Pour atteindre le nombre minimum d'observateurs, il sera possible d'effectuer la même expérience à plusieurs reprises au même endroit ou dans plusieurs endroits. Les notes obtenues dans les différents lieux utilisés pour une session d'observation par des spécialistes pourront être traitées ensemble sur le plan statistique.

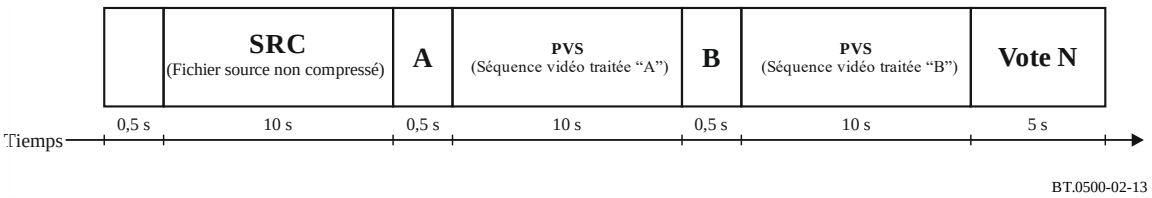
A8-3 Cellule test de base

La séquence qui sera présentée aux spécialistes devrait être organisée de manière à créer une cellule test de base (BTC) pour chaque ensemble de conditions de codage à évaluer (voir la Fig. 2-13).

Les extraits de la séquence source de référence (SRC) et des séquences vidéo traitées (PVS) à examiner dans une cellule BTC devraient toujours être tirés de la même séquence vidéo, afin que les spécialistes soient en mesure de repérer toute amélioration de la qualité visuelle offerte par les algorithmes de compression testés.

FIGURE 2-13

Organisation d'une cellule test de base pour le protocole d'observation par des spécialistes



BT.0500-02-13

La cellule BTC devrait être organisée comme suit:

– écran mis à gris moyen (valeur moyenne sur l'échelle de luminance) pendant 0,5 seconde;
– présentation de 10 secondes de l'extrait vidéo de référence non compressé;
– affichage pendant 0,5 seconde du message «A» (première vidéo à évaluer) sur fond gris moyen;
– présentation de 10 secondes d'une version altérée de l'extrait vidéo;
– affichage pendant 0,5 seconde du message «B» (deuxième vidéo à évaluer) sur fond gris moyen;
– présentation de 10 secondes d'une version altérée de l'extrait vidéo;
– affichage pendant 5 secondes d'un message demandant aux observateurs de donner leur avis.

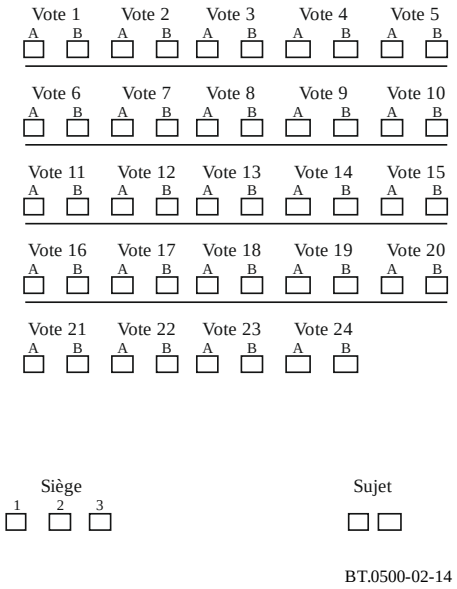
Le message «Vote» devrait être suivi d'un numéro facilitant le report sur la feuille de notation.

A8-4 Feuille de notation et échelle de notation

Comme le montre la Fig. 2-13, la présentation des extraits vidéo devraient être organisée de telle sorte que la séquence de référence non altérée (séquence SRC) soit diffusée en premier et suivie des deux séquences vidéo altérées (séquence PVS). L'ordre de présentation des séquences PVS devrait être modifié de manière aléatoire pour chaque cellule BTC et les observateurs ne devraient pas connaître l'ordre de présentation.

FIGURE 2-14

Exemple de feuille de notation pour une session d'observation par des spécialistes avec 24 cellules BTC



Une échelle numérique comprenant 11 niveaux allant de 10 (dégradation imperceptible) à 0 (dégradation très gênante) est utilisée.

Le Tableau 2-4 donne des indications sur la signification des 11 niveaux de l'échelle numérique.

TABLEAU 2-4

Signification des 11 niveaux de l'échelle numérique

Note	Dégradation	
10	Imperceptible	
9	Légèrement perceptible	À un endroit
8		Partout
7	Perceptible	À un endroit
6		Partout
5	Clairement perceptible	À un endroit
4		Partout
3	Gênante	À un endroit
2		Partout
1	Extrêmement gênante	À un endroit
0		Partout

Il est demandé aux observateurs de remplir un questionnaire comprenant deux cases («A» et «B») pour chaque cellule BTC, en indiquant dans chacune de ces deux cases une note choisie selon l'échelle numérique à 11 niveaux.

La Figure 2-14 donne un exemple de feuille de notation pour une session comprenant 24 cellules BTC.

Pour chaque cellule BTC, les observateurs remplissent la case identifiée par la lettre **A** (pour noter l'extrait vidéo apparaissant en premier) et la case identifiée par la lettre **B** (pour noter l'extrait vidéo apparaissant en deuxième).

La présentation de l'extrait vidéo original non altéré permet aux spécialistes d'évaluer plus facilement les éventuelles dégradations.

La signification des 11 niveaux de l'échelle numérique devrait être expliquée en détail lors de «sessions de formation», comme indiqué ci-après.

A8-5 Conception du test et création d'une session

L'ordre de présentation des cellules BTC devrait être fixé de manière aléatoire par le concepteur du test, de telle sorte que le même extrait vidéo ou le même extrait altéré ne soit pas présenté deux fois de suite.

Toutes les sessions d'observation devraient commencer par une «phase de stabilisation» comprenant la «meilleure» cellule BTC, la pire cellule BTC et deux cellules BTC de «qualité moyenne» figurant dans chaque session de test. Les observateurs auront ainsi immédiatement un aperçu des différentes qualités, dès le début de la session.

Si la session d'observation dure plus de 20 minutes, le concepteur du test devrait la scinder en deux (ou plus) sessions d'observation séparées, ne dépassant pas 20 minutes chacune. Dans ce cas, il devrait y avoir une «phase de stabilisation» avant chaque session.

A8-6 Formation

Même si cette procédure s'adresse à des spécialistes, il est préférable d'organiser une courte session d'observation de formation (5 à 6 cellules BTC) avant chaque expérience.

Les séquences vidéo utilisées pour la session de formation pourront être les mêmes que celles utilisées pour les véritables sessions, mais l'ordre de présentation devrait être différent.

Il convient de former les observateurs à l'utilisation de l'échelle à 11 niveaux en leur demandant de regarder attentivement les extraits vidéo diffusés immédiatement après les messages «A» et «B» sur l'écran et de déterminer s'ils peuvent voir une différence par rapport à l'extrait vidéo diffusé en premier (la séquence SRC).

A8-7 Collecte et traitement des données

Les notes devraient être collectées à la fin de chaque session et saisies dans un tableur pour calculer les valeurs MOYENNES.

Il est souhaitable d'effectuer une sélection a posteriori des observateurs moyennant une corrélation linéaire de Pearson.

Cette fonction de corrélation devrait être appliquée compte tenu de toutes les notes de chaque sujet par rapport aux notes moyennes d'opinion (MOS); un seuil pourra être fixé afin de définir pour chaque observateur s'il est «acceptable» ou «refusé» (la Recommandation UIT-T P.913 propose d'utiliser une valeur seuil de «refus» égale à 0,75).

A8-8 Conditions d'utilisation des résultats obtenus avec le protocole d'observation par des spécialistes

Le protocole d'observation par des spécialistes (EVP) pourra être utilisé lorsque le temps et les ressources disponibles ne permettent pas de mener à bien une expérience d'évaluation subjective formelle.

Une évaluation EVP prend moins de temps qu'une évaluation subjective formelle et peut être menée dans un environnement «informel», pour autant que cet environnement soit protégé de toute perturbation visuelle et sonore extérieure.

Les seules conditions à respecter impérativement concernent l'éclairage ambiant et les conditions d'observation (écran, angle et distance d'observation) comme indiqué ci-dessus.

A8-9 Limites de l'utilisation des résultats des tests EVP

Bien qu'il soit établi que le protocole EVP peut fournir des résultats acceptables avec seulement neuf observateurs, on ne peut pas considérer que les notes moyennes d'opinion obtenues avec une expérience EVP remplacent les résultats que l'on peut obtenir avec une expérience d'évaluation subjective formelle.

Les notes moyennes d'opinion obtenues avec le protocole EVP peuvent être utilisées pour avoir une première indication du niveau de dégradation.

Les notes moyennes d'opinion obtenues avec le protocole EVP peuvent être utilisées pour faire un premier classement des mécanismes de traitement vidéo testés.

Lorsqu'on le juge pratique ou nécessaire, une expérience EVP peut se dérouler à plusieurs endroits en parallèle, pour autant que les conditions et la distance d'observation ainsi que la conception du test soient identiques.

Si le nombre d'observateurs spécialistes participant à la même expérience EVP, qu'elle se déroule dans un seul ou dans plusieurs endroits, est égal ou supérieur à 15, les données subjectives brutes peuvent être traitées pour obtenir une note moyenne d'opinion, l'écart type et l'intervalle de confiance, qui pourront aider à classer de manière plus précise les cas testés. Dans ce dernier cas, une analyse statistique déductive plus précise pourra être réalisée, par exemple un test T de Student.

Pièce jointe 1
(pour information)
à l'Annexe 8
de la Partie 2

Application du protocole d'observation par des spécialistes et incidences liées à la présence d'un grand nombre d'observateurs spécialistes

La présente pièce jointe informative fournit des informations sur les résultats de deux sessions différentes d'évaluation subjective de séquences vidéo HD et UHD codées, organisées conformément au protocole EVP lors de la 117ème réunion du Groupe d'expert pour les images animées (MPEG), en application des dispositions de l'Annexe 8, afin de classer de manière rapide et fiable deux méthodes différentes de codage de la source.

Étant donné qu'un grand nombre de spécialistes ont participé à la 117ème réunion du Groupe MPEG, le nombre d'observateurs participant aux deux sessions EVP était largement supérieur au nombre de neuf observateurs préconisé dans l'Annexe 8 de la Partie 2 de la présente Recommandation; 30 spécialistes ont participé à la session de test EVP HD et 32 spécialistes ont participé à la session de test EVP UHD.

Grâce à la large participation d'observateurs spécialistes, on a pu analyser les notes moyennes d'opinion, afin de vérifier le niveau de fiabilité lié à l'utilisation de l'Annexe 8 lorsqu'il s'agit de classer des séquences vidéo codées.

Dans le cadre de cette évaluation, on considère quatre groupes d'observateurs (composés respectivement de 9, 12, 15 et 18 observateurs) et on effectue une comparaison entre les valeurs des notes moyennes d'opinion obtenues à partir du groupe de neuf spécialistes et celles obtenues à partir des groupes de 12, 15 et 18 observateurs.

L'objectif consistait à comparer le classement obtenu avec neuf spécialistes (protocole EVP) et les classements obtenus avec 12, 15 et 18 spécialistes (expérience d'évaluation subjective formelle).

Il ressort de la Fig. 2-15 (expérience portant sur un contenu UHD) et de la Fig. 2-16 (expérience portant sur un contenu HD) que les résultats des classements obtenus dans les quatre cas considérés sont très similaires.

Si l'on prend comme référence les résultats obtenus avec le groupe de 18 observateurs, on peut tracer les graphiques des Figures 2-15 et 2-16 en classant les séquences évaluées selon les valeurs des notes moyennes d'opinion obtenues avec le groupe de 18 observateurs (courbe continue rouge).

Les autres courbes des graphiques représentent les résultats obtenus avec le groupe de neuf observateurs (courbe en pointillés rouge), de 12 observateurs (courbe en traits discontinus bleue) et de 15 observateurs (courbe continue verte).

En observant les résultats représentés dans les Figures 2-15 et 2-16, il convient de noter que:

– les courbes représentant les résultats obtenus avec les groupes de 15 et de 18 observateurs présentent une pente homogène, concernant aussi bien les notes moyennes d'opinion pour les séquences de bonne qualité que celles pour les séquences de mauvaise qualité;

– les courbes représentant les résultats obtenus avec les groupes de neuf et de 12 observateurs présentent certaines «inversions» dans le classement par rapport à la courbe correspondant au groupe de 18 observateurs, bien que ces variations aient une portée limitée.

En conclusion, les expériences EVP décrites ci-avant témoignent de la très bonne qualité des résultats obtenus au moyen du protocole EVP, ce qui vient confirmer les explications données dans l'Annexe 8, à savoir que, même si l'on ne peut pas considérer qu'il remplace en tout point une expérience d'évaluation subjective formelle, le protocole EVP pourrait être considéré comme une procédure d'évaluation stable, qui fournit des résultats très proches de ceux obtenus lorsqu'un nombre beaucoup plus important d'observateurs est disponible et qu'une évaluation subjective formelle est réalisée.

FIGURE 2-15

Classement pour l'expérience UHD en fonction du nombre d'observateurs

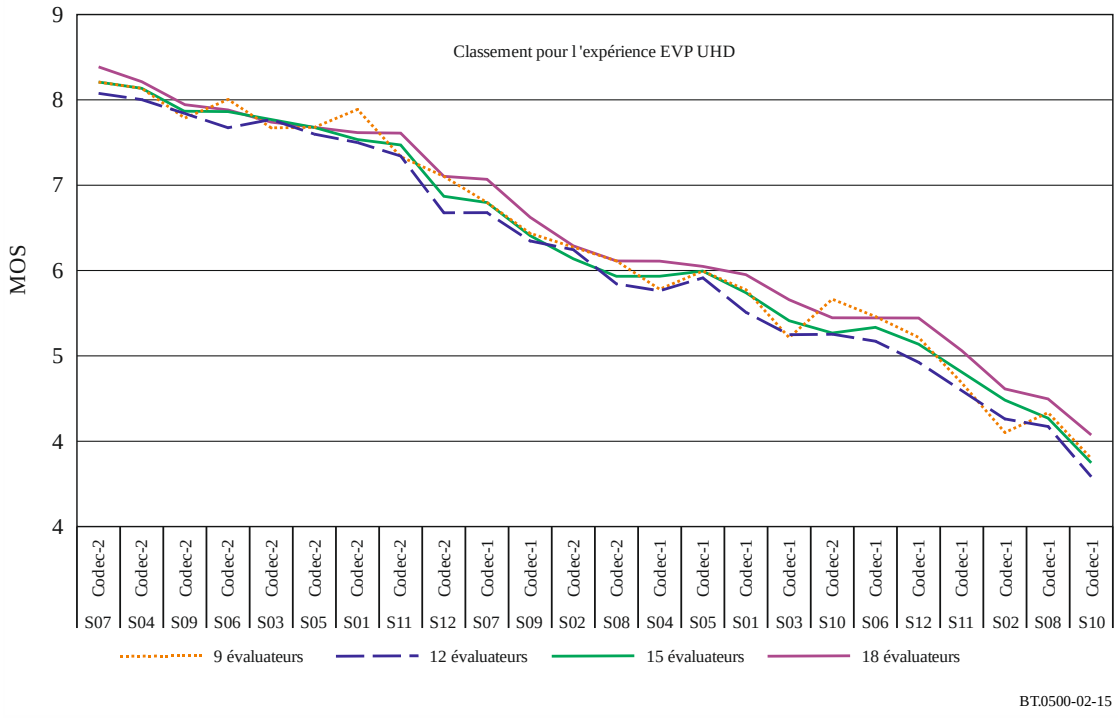
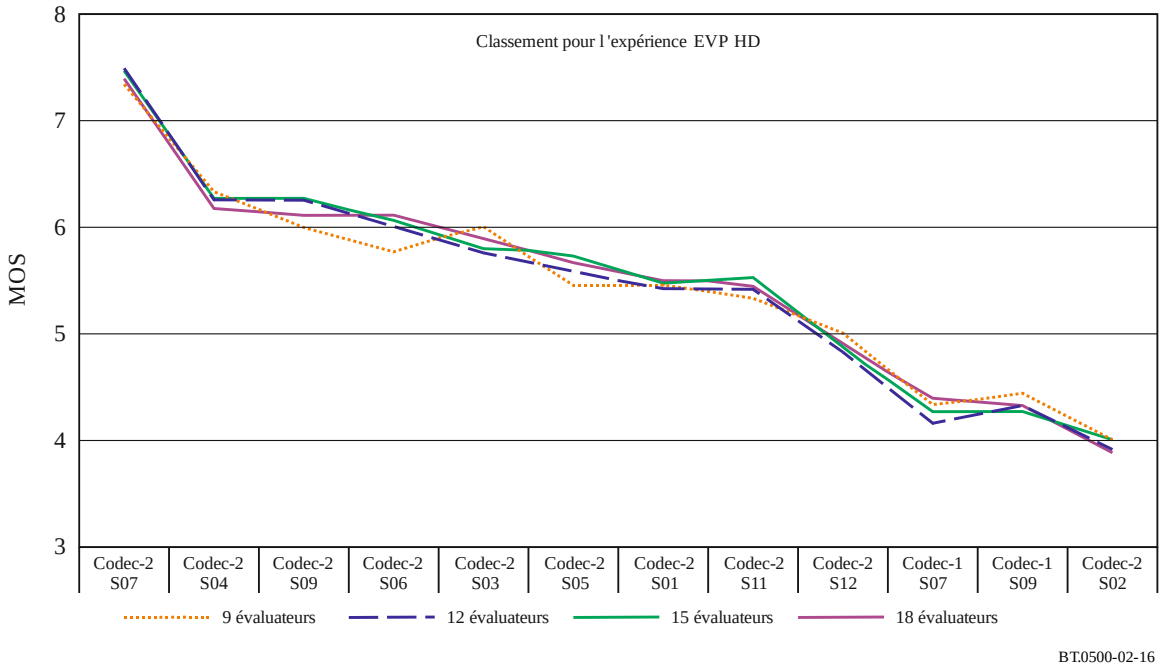


FIGURE 2-16

Classement pour l'expérience HD en fonction du nombre d'observateurs



PARTIE 3

Méthodologies d'évaluation subjective de la qualité des images pour les applications

La conception d'un essai d'évaluation subjective devrait tenir compte de considérations propres aux applications. La Partie 3 donne des orientations concernant l'évaluation subjective de la qualité des images pour différents formats d'images et applications.

Annexe 1 de la Partie 3

A1-1 Introduction

La présente annexe, qui est destinée à être utilisée conjointement avec les Parties 1 et 2 de la présente Recommandation, fournit des précisions sur l'application des méthodes générales que décrit la Recommandation à l'évaluation subjective des systèmes numériques de qualité égale ou presque équivalente à celle des systèmes de télévision classiques. Les éléments de procédure présentés ici, ainsi que les informations générales correspondantes, concernent les essais de codecs (ou de systèmes) servant à transmettre des images produites conformément à la Recommandation UIT-R BT.601 dans les applications de contribution et de distribution ainsi que d'émission.

Pour les applications de distribution, la qualité peut être définie par rapport au jugement subjectif d'observateurs. Des codecs de ce type peuvent donc théoriquement être évalués subjectivement en fonction des spécifications de qualité ainsi établies. La qualité d'un codec destiné à des applications de contribution ne pourrait cependant pas en théorie, être définie en termes de caractéristiques subjectives, car le signal qu'il produit n'est pas destiné au visionnage immédiat, mais à la postproduction, à l'enregistrement et/ou au codage pour transmission ultérieure. Du fait qu'il est difficile de définir cette qualité pour une série d'opérations de postproduction différentes, le principe choisi a consisté à définir les performances d'une chaîne d'appareils, y compris pour la fonction de postproduction, considérée comme représentative d'une application de contribution réelle. Cette chaîne pourrait normalement être constituée d'un codec, puis d'une fonction de postproduction studio (ou d'un autre codec dans le cas de l'évaluation de la qualité intrinsèque de la contribution), puis d'un autre codec avant présentation du signal à l'observateur. L'adoption de cette stratégie pour la spécification des codecs destinés aux applications de contribution a pour conséquence que les

procédures de mesure décrites dans la présente Recommandation peuvent aussi être utilisées pour les évaluer.

En matière d'évaluations subjectives, pour lesquelles on dispose d'une expérience considérable, on peut faire des recommandations sur les conditions expérimentales et sur la méthodologie. Il ne faut cependant pas oublier, lorsqu'on fixe des objectifs de qualité ou de dégradation, que les méthodes existantes peuvent donner non des notes subjectives absolues, mais des résultats partiellement affectés par les conditions de référence ou d'ancrage choisies. On peut adopter les mêmes méthodologies pour des codecs à longueur de mot fixe ou variable, inter ou intratrames, mais leur nature pourra influencer le choix des séquences d'essai.

La méthode la plus fiable pour le classement de codecs de haute qualité consiste à évaluer tous les systèmes en présence simultanément et dans des conditions identiques. Lorsqu'il n'existe que de très légères différences de qualité, les essais réalisés indépendamment ne peuvent donner qu'une indication et non une preuve indiscutable de supériorité.

Il pourrait être utile d'évaluer subjectivement les dégradations en fonction du taux d'erreur binaire sur la liaison entre le codeur et le décodeur. On ne possède actuellement qu'une connaissance expérimentale des statistiques d'erreurs réelles insuffisante pour pouvoir recommander des paramètres d'un modèle rendant compte du groupement des erreurs ou des salves d'erreurs. Tant que l'on ne disposera pas de données suffisantes, la loi de Poisson pourra être appliquée aux erreurs.

A1-2 Conditions d'observation

Pour les évaluations subjectives, les conditions générales d'observation sont celles du § 2 de la Partie 1. Les conditions d'observation propres aux évaluations subjectives des systèmes numériques sont indiquées dans les paragraphes qui suivent.

A1-2.1 En laboratoire

Le laboratoire a pour objet de fournir les conditions critiques propices à l'examen des systèmes. Les conditions d'observation propres aux évaluations subjectives en laboratoire sont indiquées dans le Tableau 3-1 qui suit.

TABLEAU 3-1

Conditions d'observation propres aux évaluations subjectives des systèmes numériques en laboratoire

Condition	Élément	Valeurs
a	Rapport de la distance d'observation à la hauteur de l'image	4 *H* et 6 *H* [1)]
b	Luminance de crête de l'écran	70 cd/m^2
c	Angle d'observation sous-tendu par la zone d'arrière-plan qui répond aux spécifications	≥ 43° (hauteur) × 57° (largeur)
d	Visualisation	Haute qualité, taille d'écran d'au moins 50 cm (20 pouces) [2)]

1) 6 *H* est la distance préférée pour l'évaluation de systèmes de télévision à définition standard numérique; toutefois, on peut également recourir à des observateurs situés à une distance de 4 *H* à condition que les résultats soient consignés séparément.

2) Il est prouvé que la taille d'écran peut influer sur les résultats des évaluations subjectives: les expérimentateurs sont donc priés d'indiquer explicitement la taille d'écran ainsi que la marque et le modèle des écrans qu'ils utilisent dans leurs expériences.

A1-2.2 À domicile

Cet environnement est censé offrir un moyen d'évaluer la qualité de la chaîne TV numérique du point de vue du consommateur. Les conditions d'observation propres aux évaluations subjectives des systèmes de télévision à définition standard à domicile sont indiquées dans le Tableau 3-2.

TABLEAU 3-2

Conditions d'observation propres aux évaluations subjectives des systèmes numériques à domicile

Condition	Élément	Valeurs
a	Rapport de la distance d'observation à la hauteur de l'image	6 *H*
b	Taille de l'écran pour un format 4/3	De 25 à 29 pouces[1)]
c	Taille de l'écran pour un format 16/9	De 32 à 36 pouces[1)]
d	Écran standard	TV à définition standard
e	Luminance de crête de l'écran	200 cd/m^2
f	Éclairement de l'environnement sur l'écran (la lumière incidente de l'environnement qui arrive sur l'écran doit être mesurée perpendiculairement sur l'écran)	200 Lux

1) Cet écran satisfait aux règles de la distance préférée d'observation; celle-ci est égale à 6 *H*.

A1-3 Méthodes d'évaluation

A1-3.1 Évaluation de la qualité intrinsèque de l'image

Lorsqu'on évalue un codec destiné à la distribution, cette qualité est celle d'images décodées après un seul passage dans une paire de codecs. Pour ceux affectés aux contributions, la qualité intrinsèque peut être évaluée à la sortie de plusieurs codecs en cascade, pour simuler des conditions d'exploitation types.

Lorsque la plage de qualité à évaluer est étroite, comme ce sera normalement le cas pour des codecs de télévision, la méthodologie à utiliser est celle de la variante II de la méthode double stimulus utilisant une échelle de qualité continue décrite dans la présente Recommandation. La séquence de source initiale sera utilisée comme référence. La durée des séquences de présentation est encore à l'étude. Dans de récentes expériences sur des codecs destinés à la vidéo en composantes 4:2:2, il a été jugé avantageux de modifier la présentation par rapport à celle qui est décrite dans la présente Recommandation. Des images composites ont été utilisées comme référence supplémentaire pour obtenir un niveau de qualité plus faible permettant d'évaluer la qualité du codec.

Il est recommandé d'utiliser dans le cadre de l'évaluation au moins six séquences d'images, plus une destinée à l'entraînement avant le début de l'expérience. Les séquences doivent être comprises entre «moyennement critique» et «critique» pour l'application à débit binaire réduit considérée.

D'un bout à l'autre de la présente annexe, on insiste sur l'importance qu'il y a de tester des codecs numériques au moyen de séquences d'images critiques pour la réduction de débit binaire en télévision. On peut donc raisonnablement se demander quel est le niveau critique d'une séquence donnée pour une application déterminée de la réduction du débit binaire ou si une séquence est plus critique qu'une autre. Une réponse simple, mais pas particulièrement utile, est de dire que la notion de niveau critique revêt une signification très différente selon les codecs. C'est ainsi qu'une image fixe très détaillée pourrait bien être critique pour un codec intratrame, tandis que pour un système intertrames capable

d'exploiter les similitudes entre images, la même scène ne poserait aucun problème. Certains types de séquences où figurent des textures mouvantes et des mouvements complexes seront critiques pour toutes les catégories de codecs et ce sont donc eux qu'il est le plus utile de produire ou de reconnaître. Les mouvements complexes peuvent être d'une forme prévisible par l'observateur, mais non par les algorithmes de codage; c'est le cas des déplacements tortueux périodiques.

Un examen de mesures statistiques possibles du niveau critique de l'image (obtenues par exemple au moyen de méthodes corrélatives, spectrales, entropiques conditionnelles, etc.) a permis de découvrir une mesure simple, mais utile basée sur une mesure de l'entropie adaptative intertrames-intratrame. Cette méthode a été utilisée pour «calibrer» les séquences d'images que l'UIT-R propose d'employer dans les essais de codecs à 34, 45 et 140 Mbit/s et s'est révélée utile dans le choix de celles à adopter. La manière la plus facile de réaliser ces mesures sur les séquences est de transférer celles-ci à des ordinateurs de traitement de l'image, puis de les soumettre à un logiciel d'analyse.

Lorsqu'il n'est pas possible de recourir à ces techniques, on pourra utiliser les indications générales qui suivent sur la manière de choisir des images critiques.

a) *Codecs intratrame à mots de longueur fixe*

S'il est possible et justifié d'évaluer ces codecs sur des images fixes, il est recommandé d'employer des séquences mobiles car le résultat du bruit de codage est plus facile à observer et cette solution est plus représentative de la télévision réelle. Si l'on emploie des images fixes dans la simulation de codecs sur ordinateur, le traitement doit être effectué sur la totalité de la séquence d'évaluation afin de conserver l'aspect temporel de tout bruit à la source, par exemple. Les scènes choisies doivent contenir le plus possible des détails suivants: régions texturées fixes et mobiles (certaines colorées), objets mobiles et immobiles avec des arêtes aiguës fortement contrastées (quelques-unes en couleur) dans diverses orientations, zones uniformes fixes d'un gris moyen. Au moins une des séquences de l'ensemble doit présenter un bruit de source à peine perceptible et une, au minimum, doit être artificielle, c'est-à-dire produite par ordinateur pour s'affranchir des imperfections des caméras, comme celles dues à l'ouverture d'analyse et au traînage.

b) *Codecs intertrames à mots de longueur fixe*

Les scènes choisies doivent toutes comporter des mouvements et le plus possible des détails suivants: régions texturées mobiles (certaines en couleur), objets comportant des arêtes et fortement contrastés (certains colorés) se déplaçant dans diverses orientations, perpendiculaires aux arêtes. Une séquence au moins dans l'ensemble doit comporter un bruit à la source à peine perceptible et une autre (au minimum) doit être artificielle.

c) *Codecs intratrame à mots de longueur variable*

Il est recommandé d'essayer ces codecs avec des séquences d'images mobiles pour les mêmes raisons que dans le cas des codecs à longueur de mot fixe. Il est à noter que, par leur codage à longueur de mot variable et leur mémoire tampon, ces codecs peuvent répartir dans toute l'image la capacité binaire disponible pour le codage. C'est ainsi que si la moitié d'une image est constituée d'un ciel monotone, que l'on peut coder avec une faible capacité, on en économise pour le reste de l'image qui peut alors être reproduite avec une grande qualité, même si elle est critique. La conclusion importante à tirer est que, si une séquence d'images doit être critique pour un codec de ce type, toutes les parties de l'écran doivent comporter des détails avec des textures fixes et en mouvement et autant de couleurs qu'il est possible; il doit aussi y figurer des objets à contours fins et très contrastés. Au moins une séquence de l'ensemble doit présenter un bruit à la source à peine perceptible et une (au minimum) doit être artificielle.

d) *Codecs intertrames à mots de longueur variable*

Il s'agit là de la catégorie de codecs la plus élaborée et du genre qui exige les images les plus difficiles pour les pousser à leurs limites. Il faut non seulement que toutes les parties de l'image comportent des détails comme pour les codecs intratrame à mots de longueur variable, mais aussi que ces détails soient en mouvement. De plus, comme de nombreux codecs utilisent des méthodes de compensation des mouvements, les déplacements pendant la séquence doivent être complexes. On peut citer les exemples suivants: scènes avec changements de cadrage et de focales simultanés, scènes ayant comme fond un rideau structuré ou comportant des détails, agité par le vent, scènes avec des objets tournant dans les trois dimensions de l'espace, scènes où des objets détaillés accélèrent leurs mouvements de traversée de l'écran. Toutes les scènes doivent comporter d'importants mouvements d'objets à des vitesses différentes, des textures et des arêtes très contrastées, ainsi que des couleurs variées. Au moins une séquence de l'ensemble doit présenter un bruit de source à peine perceptible et une autre (au minimum) doit montrer des mouvements de caméra complexes synthétisés sur ordinateur à partir d'une image fixe naturelle, de manière qu'elle soit exempte de défauts dus au bruit et au traînage de la caméra. Enfin, une séquence au moins doit être entièrement produite sur ordinateur.

A1-3.2 Évaluation de la qualité de l'image après post-traitement

Le but de cette évaluation est de juger si un codec destiné à des applications de contribution est bien adapté à cette fonction du point de vue d'opérations de post-traitement particulières telles qu'incrustations, ralenti ou recadrage électronique. Pour cette évaluation, le matériel employé doit au minimum assurer un passage dans le codec étudié, suivi par le post-traitement considéré, puis par l'observation. Il peut toutefois être plus représentatif des applications de contribution d'insérer d'autres codecs après le post-traitement.

La méthodologie à utiliser est celle de la variante II de la méthode à double stimulus utilisant une échelle de qualité continue. Dans ce cas, la référence sera cependant la séquence à la source soumise au même post-traitement que les images décodées. Si l'on juge avantageux de prévoir une référence de moindre qualité, cette dernière doit aussi être soumise au même post-traitement.

Les séquences d'essai nécessaires à l'évaluation du post-traitement sont soumises à des critères de criticité exactement identiques à ceux des séquences des autres applications numériques. Il peut cependant être difficile de les respecter dans les scènes d'avant-plan pour les incrustations, car elles comportent généralement d'importantes parties d'un bleu uniforme.

Du fait des contraintes matérielles que comporte la possibilité de devoir évaluer un codec avec plusieurs opérations de post-traitement, le nombre des séquences d'images utilisées peut être au minimum de trois, avec une supplémentaire pour la démonstration. La nature des séquences dépendra de l'opération de post-traitement étudiée, mais devra être comprise entre «modérément critique» et «critique» pour la télévision à débit binaire réduit et pour le procédé en question. Pour évaluer le ralenti, une vitesse de reproduction égale au 1/10 de celle de la source peut convenir.

A1-3.3 Évaluation du comportement en présence de défauts

Quand on évalue subjectivement les défauts des images des codecs imputables à des imperfections dans le canal de transmission ou d'émission, il faut sélectionner au moins cinq taux d'erreur binaires ou situations de transmission ou d'émission données, mais de préférence davantage, répartis de façon à peu près logarithmique et couvrant bien toute la gamme où se produisent les dégradations dues au codec et comprise entre «imperceptible» et «très gênant».

Il est possible que l'on doive évaluer des codecs à des taux d'erreur binaires de transmission provoquant des distorsions passagères visibles si rares qu'il risque de ne pas s'en produire pendant la durée d'une séquence de 10 s. Il est alors évident que le rythme de présentation proposé ici ne convient pas.

Si l'on enregistre l'image sortant d'un codec avec un taux d'erreur binaire assez faible (donnant un petit nombre de distorsions passagères visibles pendant une période de 10 s) en vue d'un montage pour constituer une séquence d'évaluation subjective, il faut veiller à ce que l'enregistrement utilisé soit représentatif de l'image sortant du codec observé pendant une période de temps plus longue.

Comme il faut étudier la qualité du codec pour une série de taux d'erreur binaires de transmission, trois séquences, plus une pour la démonstration, suffiront probablement compte tenu des contraintes matérielles. La durée d'une séquence doit être d'environ 10 s, mais il convient de noter que les observateurs préféreront peut-être une durée de 15 à 30 s. Cette séquence doit être comprise entre «moyennement critique» et «critique» pour la télévision à débit binaire réduit.

Comme les essais couvriront toute la plage des dégradations, la méthode à double stimulus utilisant une échelle de dégradation est applicable et doit être utilisée.

A1-3.4 Caractéristiques de défaillance fonction du contenu de l'image

Le concept général de caractéristiques de défaillance fonction du contenu de l'image est donné dans l'Annexe 1 de la Partie 1. Pour appliquer ce concept aux systèmes de télévision numérique à définition normale, il faut procéder de la manière suivante.

A1-3.4.1 Définition de la criticité

Il faut définir une certaine mesure appelée «criticité», qui représente les caractéristiques du système de télévision numérique sous test et qui est déterminée par une mesure objective. Le système de télévision numérique servant d'exemple est le système MPEG-2 MP@ML et on applique la méthode – à quantificateur fixe – de détermination de la criticité basée sur l'entropie, qui est décrite dans la Recommandation UIT-R BT.1210.

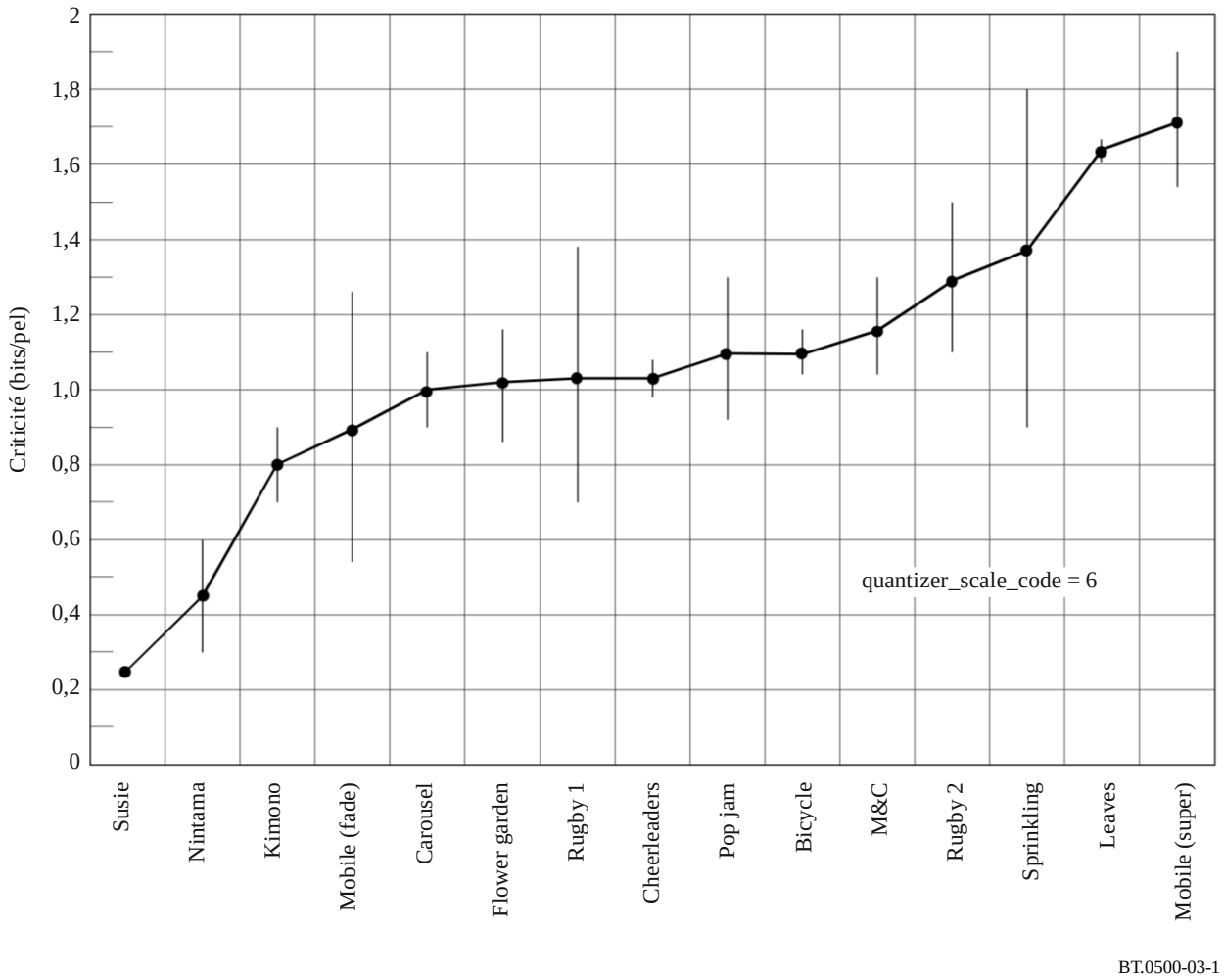
A1-3.4.2 Procédure d'établissement des caractéristiques de défaillance fonction du contenu de l'image

– *Étape 1:* Mesurer la criticité des séquences d'essai utilisées pour l'évaluation subjective.

On mesure la criticité des séquences d'essai utilisées pour l'évaluation subjective décrite à l'étape 3 ci-dessous. La Figure 3-1 représente la moyenne et l'écart type de chaque séquence pour le système utilisé comme exemple. La criticité de la plupart des séquences est comprise entre 0,8 et 1,4 bits/pixel. L'écart type de certaines séquences est grand car le contenu de l'image varie beaucoup pendant la séquence.

FIGURE 3-1

Moyennes et écarts types de la criticité associée aux séquences d'essai

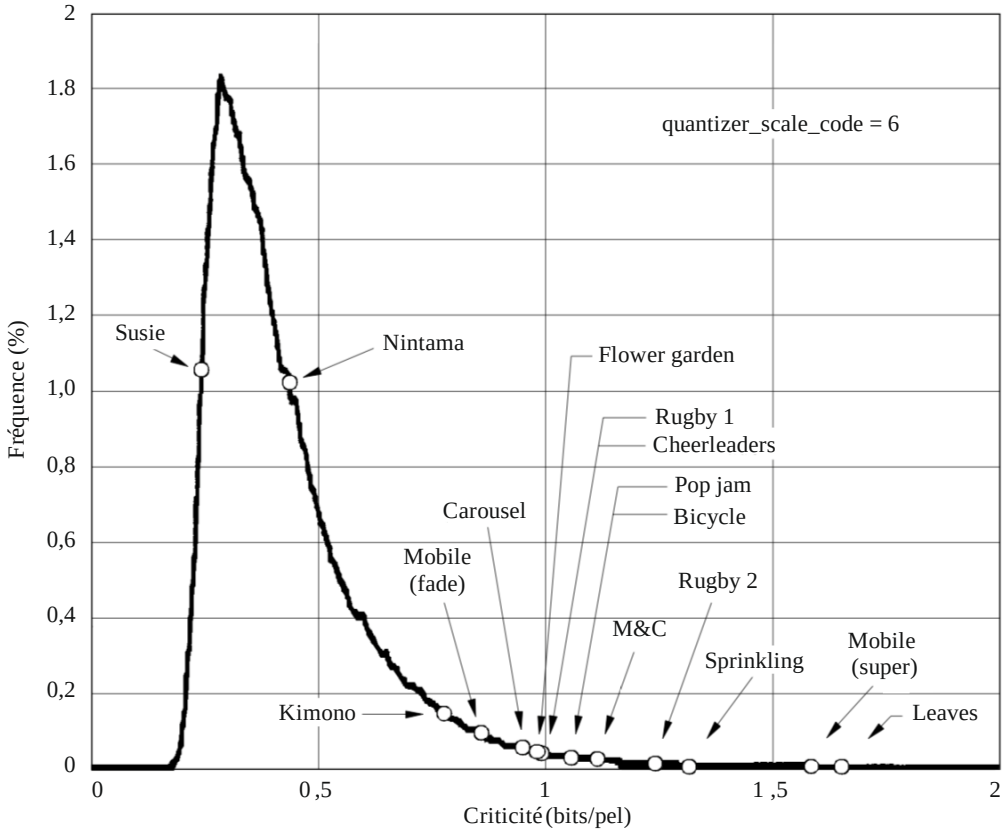


– *Étape 2:* Mesurer, sur une longue période temporelle, la distribution de la criticité associée à des programmes de radiodiffusion.

La distribution de la criticité associée à des programmes de radiodiffusion télévisuelle est mesurée sur une période temporelle suffisamment longue, par exemple une semaine. La Figure 3-2 montre un exemple de distribution mesurée sur une semaine, soit un total de 130 h pour des signaux de radiodiffusion NTSC, qui ont été convertis en signaux à composantes de luminance et de chrominance *(Y/C)* pour la mesure. La fréquence d'occurrence de la criticité pour les programmes de radiodiffusion télévisuelle a été calculée tous les 5×10^{-3} bits/pixel. Cette figure montre aussi la criticité des séquences d'essai utilisées pour l'évaluation subjective.

FIGURE 3-2

Distribution de la criticité associée à des programmes de radiodiffusion et criticité des séquences d'essai



BT.0500-03-2

– *Étape 3:* Faire une évaluation subjective de la qualité d'image du système sous test et en déduire une relation entre la criticité et la qualité d'image subjective.

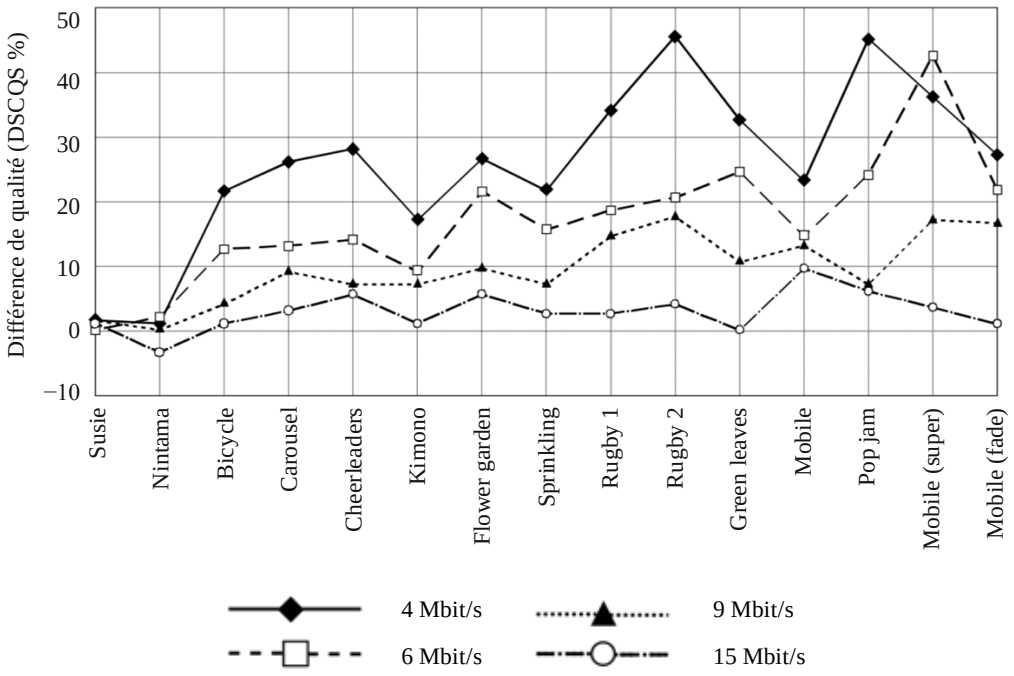
La qualité d'image du système de télévision numérique est évaluée à l'aide de la méthode à double stimulus utilisant une échelle de qualité continue (DSCQS). La combinaison du résultat de l'évaluation subjective et de la criticité obtenue à l'étape 1 permet d'établir la relation entre la criticité et les notes de l'évaluation. La Figure 3-3 montre la qualité d'image du système utilisé comme exemple pour les débits binaires suivants: 4, 6, 9 et 15 Mbit/s. La différence de qualité (DSCQS %) sur la figure représente la dégradation par rapport à la référence: une séquence en composantes 4:2:2 d'origine. La Figure 3-4 montre la relation existant entre la criticité et la différence de qualité. Dans cet exemple, on a supposé une relation linéaire entre la criticité et la qualité d'image; les droites de régression ont été établies à l'aide de la méthode des moindres carrés. La droite de régression correspondant à chaque débit binaire est représentée sur la figure. En général, on peut appliquer une relation non linéaire en fonction des résultats de l'évaluation.

– *Étape 4:* Établir les caractéristiques de défaillance en fonction du contenu de l'image (qualité en fonction de la fréquence d'occurrence) en combinant les résultats de l'étape 3 (criticité en fonction de la qualité) et de l'étape 2 (criticité en fonction de la fréquence d'occurrence).

La combinaison des résultats obtenus aux étapes 2 et 3 permet d'établir les caractéristiques de défaillance en fonction du contenu de l'image, c'est-à-dire la distribution de la qualité d'image des programmes de télévision codés numériquement. La dégradation d'image dans les programmes de radiodiffusion télévisuelle est convertie en fréquence d'occurrence cumulative. La Figure 3-5 représente les caractéristiques de défaillance en fonction du contenu pour le système utilisé comme exemple.

FIGURE 3-3

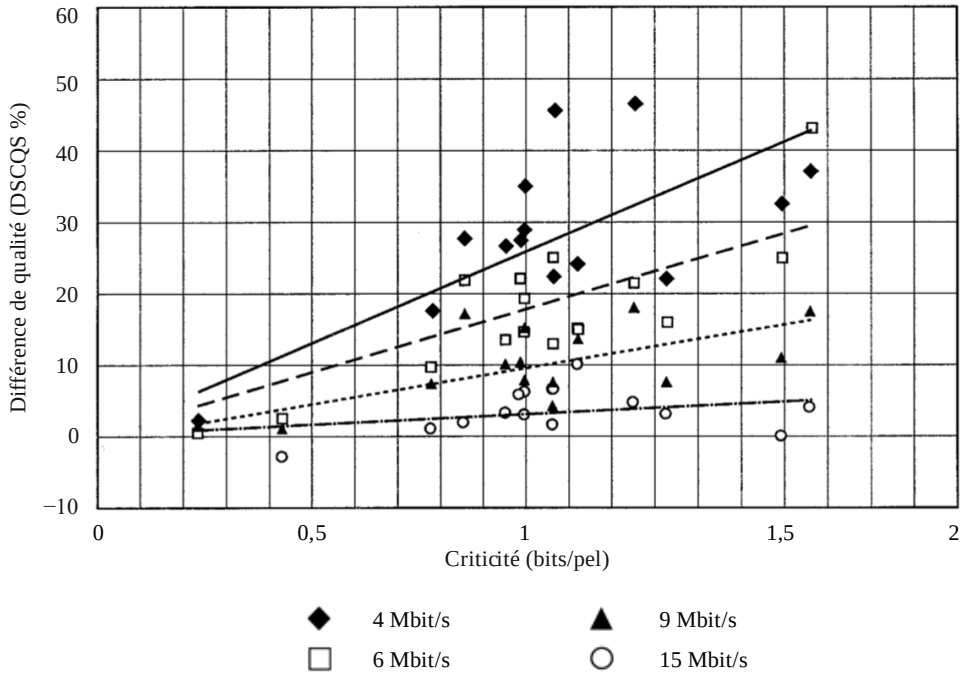
Résultat de l'évaluation subjective (système MP@ML à 6*H*)



BT.0500-03-3

FIGURE 3-4

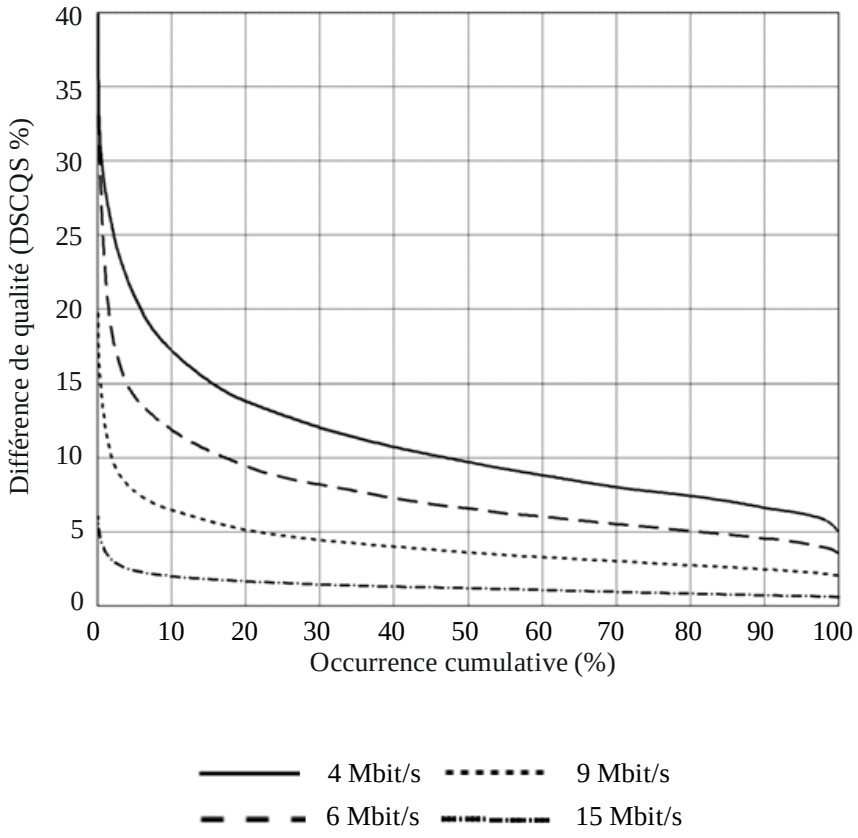
Relation entre la criticité et la note d'évaluation (MP@ML à 6*H*)



BT.0500-03-4

FIGURE 3-5

Fréquence d'occurrence cumulative de la dégradation d'image (système MP@ML à 6*H*)



BT.0500-03-5

A1-4 Notes concernant les applications

Lorsqu'il est inutile de juger la qualité ou la dégradation absolue d'un codec et qu'on ne demande que de classer les systèmes ou lorsqu'on souhaite confirmer un classement obtenu par la méthode du double stimulus, il faut utiliser une technique de comparaison entre couples de stimulus.

Telle qu'elle est décrite dans la présente Recommandation, la méthode permet une comparaison précise et donne le moyen de déterminer la relation entre couples de systèmes. Il est possible d'étendre cette méthode au classement de la qualité ou de la dégradation pour plus de deux systèmes. Selon ce principe, on établit un classement global à partir de ceux donnés par les observateurs pour tous les couples possibles de séquences d'images.

L'analyse est compliquée par le fait qu'un observateur peut juger, par exemple, que l'image A est meilleure que l'image B, elle-même meilleure que l'image C, tout en préférant C à A. On parle alors de «triade intransitive».

Un problème posé par cette méthode tient au fait que le nombre de présentations nécessaires augmente avec le carré du nombre de séquences d'essai et de codecs, si bien que la mise en œuvre de cette méthode peut se révéler impossible.

Si le canal de radiodiffusion est utilisé pour acheminer soit des trains de programmes multiples, soit des schémas de codage échelonnables ou hiérarchiques, il peut se révéler nécessaire d'adapter la méthode d'évaluation pour tenir compte de ce qui suit:

– pour savoir si un service est acceptable, le critère n'est pas forcément la transparence du codage à la source; il peut s'agir en fait de la faculté du système, pour une attribution de débit binaire donnée, d'offrir une solution de rechange convenable au service classique. Dans les essais de qualité, il convient donc peut-être de prendre pour référence des images acheminées par un système classique dans des conditions de réception normales plutôt que des images sous forme numérique non comprimée. Il convient peut-être aussi d'utiliser des éléments d'essai choisis pour être représentatifs d'une gamme de contenus de programme actuels ou futurs (voir l'Annexe 3 de la Partie 1). Au cours des essais, les conditions d'observation doivent être celles que décrivent la Partie 1 et le § A1-2 de la présente Annexe, alors que la méthode d'essai générale sera la méthode à double stimulus utilisant une échelle de qualité continue (Annexe 2 de la Partie 2); et

– une question est de savoir si le système est à même de conserver complètement les trains de programmes en cas de pleine charge du canal et de défauts de transmission. Il convient peut-être donc, au cours des essais sur les dégradations, de maintenir le canal à pleine charge et d'avoir une gamme de niveaux de dégradation représentant celle qui se produit en général dans des conditions de réception normales (voir l'Annexe 4 de la Partie 1). Au cours des essais, les conditions d'observation doivent être celles que décrivent la Partie 1 et le § A1-2 de la présente Annexe, alors que la méthode d'essai générale sera la méthode à double stimulus utilisant une échelle de dégradation (voir l'Annexe 1 de la Partie 2).

NOTE 1 – Lorsqu'on évalue des systèmes analogiques et numériques dans un même contexte, il importe de choisir un ensemble d'éléments d'essai présentant des difficultés proportionnées aux deux types de système. Dans ce cas, il peut être utile d'appliquer, en vue d'une analyse supplémentaire, la procédure d'échelonnement multidimensionnel.

Annexe 2
de la Partie 3

Méthodes d'évaluation subjective de la qualité d'image en télévision à haute définition (TVHD)

A2-1 Conditions d'observation

Sauf indication contraire dans le Tableau 3-3 ci-dessous, les conditions d'observation devraient être celles décrites au § 2 de la Partie 1.

TABLEAU 3-3

Conditions d'observation pour l'évaluation subjective de la qualité d'image en TVHD

Condition	Sujet	Valeurs
a	Rapport de la distance d'observation à la hauteur de l'image	3
b	Luminance maximale de l'écran (cd/m^2)(1)	150-250
c	Rapport de la luminance de l'écran inactif à la luminance maximale(2)	≤ 0,02
d	Rapport de la luminance de l'écran affichant seulement le niveau du noir dans une salle complètement noire à celle du blanc maximum(3)	environ 0,01
e	Rapport de la luminance de l'arrière-plan derrière l'écran à la luminance maximale de l'image	environ 0,15
f	Éclairement de la salle dû à d'autres sources(4)	faible
g	Chromaticité de l'arrière-plan	D_{65}
h	Angle sous-tendu par la zone d'arrière-plan satisfaisant les conditions ci-dessus(5). Cela doit être respecté pour tous les observateurs	53° (hauteur) × 83° (largeur)
i	Placement des observateurs	À l'intérieur d'un angle horizontal de ±30° dont le sommet est le centre de l'écran. La limite verticale est à l'étude
j	Dimensions de l'écran(6)	1,4 m (55 pouces)

(1) La luminance maximale de l'écran correspond à un signal vidéo ayant une amplitude de 100%.

(2) Il est possible que cette caractéristique soit influencée par l'éclairement de la salle et par la gamme de contraste sur l'écran.

(3) Le niveau du noir correspond à un signal ayant une amplitude de 0%.

(4) L'éclairement de la salle doit être ajusté de sorte à rendre possible les conditions c et e.

(5) Un minimum de 28° (hauteur) × 48° (largeur) est recommandé.

(6) Des valeurs ≥ 76,2 cm (30 pouces) pourront être utilisées si des écrans à la dimension spécifiée ne sont pas disponibles. Voir la Note 3 de la Partie 1.

A2-2 Méthodes d'évaluation

L'évaluation subjective de la qualité globale d'une image de TVHD fournie par un système d'émission doit être effectuée au moyen de la méthode à double stimulus avec échelle continue de qualité (Annexe 2 de la Partie 2) et en prenant l'image de qualité studio TVHD comme référence.

L'évaluation des caractéristiques de défaillance d'un système d'émission TVHD doit être effectuée au moyen de la méthode à double stimulus avec échelle de dégradation (Annexe 1 de la Partie 2), en prenant comme référence l'image du studio TVHD ou l'image d'émission non dégradée.

Lorsqu'il est question de la variation de la qualité en fonction du contenu des programmes ou des conditions de transmission rencontrées dans la pratique, on doit tenir compte des caractéristiques de dégradation composite figurant dans l'Annexe 4 de la Partie 1.

Lorsque l'on utilise ces méthodes, on doit prendre soin de mettre en évidence l'influence du format de présentation de l'image quand il n'est pas celui du système de base (par exemple, suite à conversion vers le haut). S'il y a lieu, des évaluations supplémentaires pourront être faites au moyen de différentes présentations pour tenir compte des différents formats.

Certains des systèmes d'émission de TVHD peuvent comprendre un format de télévision classique (compatibilité vers l'arrière). Il y aura donc lieu d'évaluer si, en termes de qualité d'image, les images de télévision classique insérées dans des émissions de TVHD sont adéquates. Pour ces systèmes, il convient d'appliquer les conditions d'observation et les méthodes d'évaluation spécifiées dans l'Annexe 1 de la Partie 3.

Il convient d'appliquer les concepts et procédures de base spécifiés dans l'Annexe 1 de la Partie 3 aux systèmes d'émission de TVHD numérique qui utilisent des schémas de réduction de débit binaire.

A2-3 Images d'essai

Le Rapport UIT-R BT.2245 contient une liste constituée d'une vaste gamme d'images fixes et de séquences d'images animées qu'il convient d'utiliser de préférence comme images d'essai ordinaires pour effectuer une évaluation de la qualité d'image en TVHD.

Annexe 3
de la Partie 3

Évaluation subjective de la qualité des images alphanumériques et graphiques en télétexte et dans des services similaires

Introduction

Il existe des systèmes qui traitent les images graphiques et alphanumériques et les transmettent au moyen de codes numériques appropriés. Les images alphanumériques et graphiques ont des caractères spécifiques distincts de ceux des images de télévision conventionnelle et le processus mental mis en jeu pour leur évaluation subjective peut être différent.

La présente Recommandation propose des méthodes pour évaluer la qualité subjective des images qui apparaissent dans les programmes de télévision actuels. Il est nécessaire d'étudier la qualité des images alphanumériques et graphiques qu'emploient plusieurs nouveaux services transmis dans le canal de télévision et qui utilisent des codes numériques pour décrire les images alphanumériques et graphiques. Certaines caractéristiques de transmission influencent la qualité des images affichées: la résolution de la page (nombre de lignes par page et nombre de caractères par ligne) dans le cas du codage alphamosaïque du télétexte, résolution de la matrice de caractère (nombre de pixels et de lignes par cellule) dans le cas du codage JCDR (Jeux de caractères dynamiquement redéfinissables (voir la Recommandation UIT-R BT.653)), résolution de l'image dans le cas de l'audiographie radiodiffusée, de la télécopie ou du télétexte. Il convient aussi d'étudier les effets des erreurs de

transmission qui peuvent affecter les codes. Il faut donc mesurer la qualité et déterminer des relations objectives-à-subjectives pour ces caractéristiques.

Des études ont montré que l'évaluation de la qualité de ces images nécessite des approches diverses dont les caractéristiques peuvent être différentes de celles utilisées pour les images de télévision habituelles. Des caractéristiques comme le format des pixels, la résolution des matrices de caractère, les espacements, les couleurs et la disposition influencent les différents attributs de la qualité: lisibilité, qualité, confort, gêne, effort à la lecture, fatigue et considérations esthétiques. On considère ici trois aspects essentiels: les conditions d'observation, les méthodes d'évaluation et le contexte d'évaluation.

Puisqu'il est important de définir la base des évaluations subjectives de la qualité des images alphanumériques et graphiques, tous les rapports d'essai devraient donner la description la plus complète possible des configurations d'essai, du matériel d'essai, des observateurs et des méthodes.

A3-1 Conditions de visualisation

La Partie 1 définit des conditions de visualisation pour les images de télévision qui correspondent à des niveaux d'illumination faible dans la salle. Il est vraisemblable que les images alphanumériques et graphiques pourront être regardées aussi dans les conditions normales d'éclairage. Un ensemble de conditions de visualisation supplémentaire a ainsi été suggéré pour étude: illumination de 500 lux, luminance maximale d'écran de 70 à 200 cd/m^2, rapport du contraste sur l'écran de 30 à 50 et valeur 1/4 pour le rapport de la luminance du fond (provenant des murs de la salle) à la luminance maximale de l'écran. La distance d'observation doit également être discutée (de 4 à 8 fois la hauteur de l'image).

A3-2 Méthodes d'évaluation

Un nombre considérable d'études ont été menées dans le domaine typographique. La plupart d'entre elles utilisent des «mesures de performance» comme les seuils de détection ou de reconnaissance, le taux de reconnaissance, la vitesse de lecture, etc. Très peu utilisent les «mesures subjectives» qui sont d'un usage traditionnel dans l'évaluation de la qualité des images de télévision. On pense que les nouveaux systèmes de transmission dans les canaux de télévision devront avoir de bonnes performances (par exemple, un pourcentage de bonne reconnaissance des lettres supérieur à 95%). L'échelle de qualité ou celle de dégradation de la présente Recommandation pourrait ainsi être utilisée efficacement, bien qu'il faille étudier dans quelle mesure ces échelles peuvent être reliées à la lisibilité. Une comparaison avec les méthodes d'évaluation de la qualité de la parole (UIT-T) a été tentée et une échelle à 5 notes d'«effort de lecture» a été suggérée pour la suite des études.

Une autre méthode compare les résultats d'évaluations subjectives obtenus au moyen de deux échelles à cinq notes différentes, indiquées dans le Tableau 3-4.

TABLEAU 3-4

Échelles de qualité de la lisibilité et d'effort de lecture

Échelle de qualité de la lisibilité	Échelle d'effort de lecture
Lisibilité excellente	Aucun effort de lecture
Lisibilité bonne	Attention nécessaire, mais pas d'effort de lecture appréciable
Lisibilité assez bonne	Effort de lecture modéré
Lisibilité médiocre	Effort de lecture important
Lisibilité mauvaise	Effort de lecture très important

On a trouvé important d'avoir des libellés très explicites des notes de chaque échelle. Les valeurs moyennes des notes obtenues avec l'échelle d'effort de lecture sont généralement supérieures à celles obtenues avec l'échelle de lisibilité et la dynamique utilisée par les observateurs est supérieure dans le cas de l'échelle d'effort de lecture.

Dans une autre expérience, on s'est servi de l'échelle de qualité que décrit le § A3-4.1 de la Partie 2 pour évaluer à la fois la qualité globale et la lisibilité globale d'un texte dactylographié transmis par un système de télévision à nombre de lignes et à largeur de bande variables. Dans chaque cas, on a constaté que deux modèles, l'un de complexité et de précision plus grandes, mais tous deux recourant au concept de l'addition des «échelles de dégradation» rendaient compte des effets combinés produits par des définitions horizontale et verticale limitées. On mesurait aussi la lisibilité, exprimée par la proportion de caractères correctement identifiés. Mais, dans ce cas, la lisibilité restait bonne lorsque la qualité était faible, ce qui montre qu'en général ce dernier critère est moins utile.

Une étude a comparé des méthodes de performance et des méthodes subjectives sur des textes imprimés utilisant des caractères de largeur fixe et variable. Les méthodes subjectives se sont avérées les plus sensibles. Ce même type d'étude a été renouvelé sur tube à rayons cathodiques, en utilisant cette fois uniquement les méthodes subjectives. L'utilisation de ces méthodes subjectives a permis d'obtenir des résultats concernant l'optimum visuel de la taille des matrices fixes et variables.

A3-3 Contexte d'évaluation

Une nouvelle approche pour l'évaluation des services a été proposée dans le cas où les activités des usagers du service étudié peuvent être définies de manière précise. Au lieu de présenter des images selon la méthode classique et de demander simplement une opinion subjective, les observateurs sont invités à utiliser les images qui leur sont présentées comme ils le feraient dans le cadre du service étudié et toutes les évaluations se font dans ces conditions.

Une telle émulation n'exclut pas les mesures subjectives classiques, cependant elle fournit un contexte d'évaluation subjective plus spécifique du service étudié. Elle autorise même dans certains cas, l'emploi de mesures objectives de la performance de l'observateur et la mise au point de nouvelles mesures subjectives particulièrement bien adaptées au service et aux paramètres considérés. Enfin, elle constitue une base plus fiable à partir de laquelle les évaluations faites en laboratoire peuvent être appliquées au service étudié.

Annexe 4
de la Partie 3

Évaluation subjective de la qualité d'image des services multiprogramme[5]

Introduction

Pour l'évaluation subjective de la qualité des différents programmes employant la compression et le codage à débit binaire constant (CBR) dans le cadre d'un service multiprogramme, les procédures d'évaluation subjective exposées en détail dans l'Annexe 1 ou l'Annexe 2 de la Partie 3 et la procédure décrite au § A4-2 de la présente Annexe devraient être utilisées.

[5] Y compris les services désignés par l'expression services de «multiplexage statistique» ou «Stat-Mux».

Pour l'évaluation subjective de la qualité des différents programmes employant la compression et le codage à débit binaire variable (VBR) par des méthodes telles que le multiplexage statistique ou le codage commun dans le cadre d'un service multiprogramme, les procédures d'évaluation subjective exposées en détail dans l'Annexe 1 ou l'Annexe 2 de la Partie 3 et la procédure décrite au § A4-3 de la présente Annexe devraient être utilisées.

A4-1 Informations générales sur les évaluations

- Les évaluations de la qualité des canaux classés par thème devraient être effectuées moyennant l'utilisation d'images d'essai analogues par leur contenu et leur criticité à celles qui seraient généralement transmises sur ces canaux.
- Pour évaluer la qualité de programmation globale perçue, qui varie en valeur «instantanée» sur une période donnée, les procédures décrites dans les §§ A4-2 et A4-3 devraient être utilisées.
- L'échelle de notation des résultats des systèmes utilisant des images de référence de faible qualité, selon les observations figurant dans la description de la méthode DSCQS, devrait être appliquée et examinée de manière plus approfondie en vue d'essais comparatifs entre les services multiprogramme et des images de faible qualité.

A4-2 Procédures d'évaluation subjective des images des services multiprogramme à débit binaire constant

L'évaluation subjective de la qualité des images de chaque programme de TVDN et de TVHD peut être effectuée séparément par les méthodes décrites dans l'Annexe 1 (TVDN) ou l'Annexe 2 (TVHD) de la Partie 3. Pour l'évaluation de la qualité de base du système, il convient d'utiliser la méthode d'essai générale DSCQS (décrite dans l'Annexe 2 de la Partie 2). Pour l'évaluation des programmes sujets à des dégradations de transmission, il convient d'utiliser la méthode d'essai générale DSIS (décrite dans l'Annexe 1 de la Partie 2).

A4-3 Procédures d'évaluation subjective des images des services multiprogramme à débit binaire variable

L'évaluation subjective de la qualité des images des programmes de TVDN et de TVHD avec codage VBR peut être effectuée à l'aide de la méthodologie DSCQS. Par ailleurs, le choix des images d'essai appelle une attention particulière, car la qualité d'image peut dépendre du contenu des images de tous les programmes multiplexés.

Annexe 5
de la Partie 3

Évaluation, par visionnage d'experts, de la qualité d'image des systèmes d'affichage pour l'imagerie numérique sur grand écran[6] en salle

A5-1 Introduction

Depuis quelques années, on recourt de plus en plus souvent à un visionnage d'experts pour vérifier rapidement la qualité de fonctionnement d'applications vidéo génériques.

La présente Annexe décrit une méthode d'essai qui fait intervenir des visionneurs experts et qui permettra d'obtenir des résultats cohérents d'un laboratoire à l'autre, avec un petit nombre d'observateurs experts.

A5-2 Pourquoi une nouvelle méthode faisant intervenir un «visionnage d'experts»

Il est utile à ce stade de souligner les avantages découlant de l'application de la méthode proposée.

Tout d'abord, un essai d'évaluation subjective fait nécessairement intervenir le plus souvent au moins 15 observateurs choisis parmi des personnes «non expertes», ce qui allonge la durée des essais proprement dits et impose de rechercher constamment de nouveaux observateurs. Ce nombre élevé de visionneurs est nécessaire pour obtenir la sensibilité requise afin de différencier et de classer les différents systèmes ou de les déclarer comme étant équivalents, avec le degré de confiance voulu.

En deuxième lieu, les essais traditionnels, reposant sur des observateurs non experts, peuvent ne pas révéler des différences qui pourraient devenir manifestes, même pour des personnes non expertes, au bout d'un certain temps d'observation.

Troisièmement, les méthodes d'évaluation traditionnelles consistent généralement à mesurer la qualité (ou les différences de qualité), mais non pas à identifier directement les artefacts ou les autres phénomènes physiques qui sont à la base des particularités relevées.

La méthode proposée ci-après doit apporter une solution à ces trois problèmes.

A5-3 Définition de l'expression «visionneur expert»

Dans la présente Annexe, un «visionneur expert» est par définition une personne qui connaît les équipements utilisés dans l'évaluation, qui sait «ce qu'il faut rechercher» et qui peut avoir une connaissance approfondie de la structure détaillée de l'algorithme utilisé pour traiter les programmes vidéo considérés. En tout état de cause, un «visionneur expert» est une personne ayant une longue expérience des méthodes d'évaluation de la qualité, tout en étant un professionnel du domaine spécifique. Par exemple, lorsque l'on voudra évaluer des programmes LSDI, on sélectionnera comme «visionneurs experts» des spécialistes de la production ou de la postproduction de films cinématographiques ou de programmes vidéo de haute qualité (directeurs de la photographie, spécialistes de l'étalonnage colorimétrique, etc.), et l'on tiendra compte de la capacité des candidats à formuler des jugements subjectifs précis sur la qualité de l'image et les artefacts de compression.

---

[6] L'imagerie numérique sur grand écran (LSDI, *large screen digital imagery*) est une famille de systèmes d'imagerie numérique applicables à des programmes tels que les films, pièces de théâtre, manifestations sportives, concerts, manifestations culturelles, etc., depuis la prise de vues jusqu'à la projection sur grand écran avec une haute résolution dans des cinémas, des salles ou d'autres lieux convenablement équipés.

A5-4 Sélection des évaluateurs

Dans les évaluations faisant intervenir des visionneurs experts, il s'agit de demander à des spécialistes d'évaluer la qualité de l'image visionnée et/ou la visibilité des dégradations éventuelles.

Un groupe de visionneurs experts est composé de cinq à six personnes: avec un petit nombre d'observateurs, il est plus facile de procéder à la sélection et de décider rapidement.

En fonction des besoins de l'expérience, on pourra accepter l'utilisation de plusieurs groupes de visionneurs, rassemblés en un pool (représentant par exemple plusieurs laboratoires).

Étant entendu que les visionneurs experts peuvent avoir tendance à «orienter» leur jugement lorsqu'ils évaluent leurs propres technologies, on évitera d'inclure dans les groupes des personnes ayant directement participé à la mise au point du système étudié.

Il conviendra de vérifier l'acuité visuelle, normale ou après correction, de tous les observateurs (Test de Snellen) ainsi que leur vision des couleurs (Test d'Ishihara).

A5-5 Séquences d'essai

Les séquences d'essai seront choisies en fonction de la gamme des paramètres de production et des niveaux de difficulté caractéristiques des conditions réelles dans lesquelles le système considéré sera utilisé. On choisira des séquences relativement difficiles à reproduire, mais sans excès. Dans l'idéal, on utilisera entre cinq et sept séquences d'essai.

La méthode de sélection des séquences d'essai pourra aussi dépendre de l'application en fonction de laquelle le système considéré a été mis au point.

À cet égard, les lignes qui suivent ne comporteront aucune autre indication quant à la procédure de choix des séquences d'essai: le responsable des essais choisira en fonction des éléments précités.

A5-6 Conditions d'observation

Les conditions d'observation, qui seront décrites en détail dans le rapport d'essai, respecteront les éléments indiqués dans le Tableau 3-5 et seront constantes pendant toute la durée de l'essai.

TABLEAU 3-5

Résumé des conditions d'observation

Condition d'observation	Valeur(s)	
	Minimum	**Maximum**
Dimension de l'écran (m)	6	16
Distance observateur-écran[1]	1,5 H	2 H
Luminance du projecteur (au centre de l'écran, blanc max.)	34 cd/m^2	48 cd/m^2
Luminance de l'écran (projecteur éteint)		< 1/1 000 de la luminance du projecteur

[1] On utilisera la présentation «symétrie verticale» pour une distance observateur-écran voisine de 1,5 H. Lorsque la présentation utilisée sera de type «côte à côte», la distance observateur-écran devra être de l'ordre de 2 H.

A5-7 Méthodologie

A5-7.1 Sessions d'évaluation

Une session d'évaluation (c'est-à-dire l'ensemble des séances tenues par un groupe d'observateurs) comportera deux phases (Phase I et Phase II).

A5-7.1.1 Phase I

La Phase I sera un essai subjectif formel dans des conditions contrôlées (voir le § A5-6), qui permettra d'obtenir des résultats valides, sensibles et répétitifs. Les visionneurs experts attribuent une note d'observation aux séquences qu'ils visionnent en utilisant l'échelle décrite ci-après. Les membres du groupe ne sont pas autorisés à commenter entre eux ce qu'ils voient, et n'interviennent pas dans la présentation. Pendant cette phase, les experts n'ont pas connaissance du système de codage évalué ou de l'ordre de présentation des séquences. Les séquences d'essai sont présentées en ordre aléatoire: ainsi, les observateurs ne peuvent pas être influencés dans leur jugement.

A5-7.1.1.1 Présentation des séquences

La méthode de présentation combine des éléments de la méthode d'évaluation continue à double stimulus simultané (SDSCE) décrite dans l'Annexe 6 de la Partie 2 et des éléments de la méthode de double stimulus à échelle de qualité continue (DSCQS) décrite dans l'Annexe 2 de la Partie 2. À des fins de référence, on pourra dénommer cette méthode «méthode SDS» (double stimulus simultané).

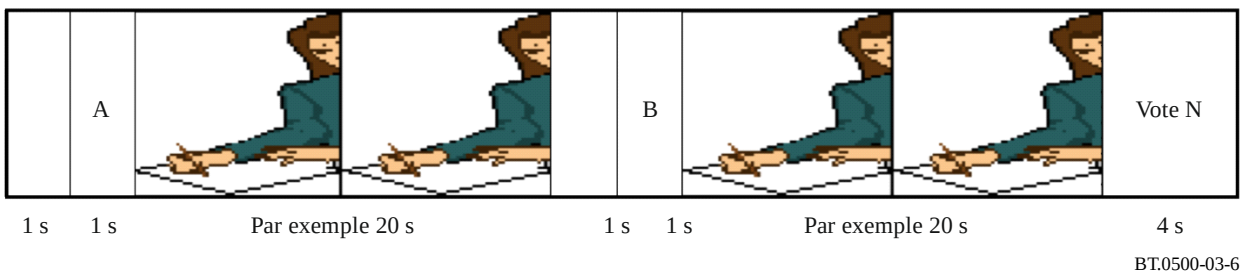
Comme avec la méthode SDSCE, on présentera chaque fois deux images sur un écran partagé. Dans la plupart des cas, l'une des sources sera la référence (image source), l'autre étant la séquence d'essai; dans d'autres cas, les deux images seront issues de la référence. La référence sera la source présentée de façon transparente (c'est-à-dire sans compression autre que la compression inhérente au support d'enregistrement). La séquence d'essai sera obtenue à partir de la source, mais traitée par l'un des systèmes évalués. Le débit et/ou le niveau de qualité correspondront aux valeurs spécifiées pour l'essai. Mais contrairement aux conditions d'application de la méthode SDSCE, les observateurs n'auront pas connaissance de la nature des deux séquences.

La présentation en écran partagé se fera soit de façon traditionnelle (sans effet miroir) soit en «symétrie verticale», par retournement horizontal de l'image occupant la partie droite de l'écran. Les images de départ étant des images plein écran, on ne pourra afficher que la moitié de chaque image, et l'on veillera donc à montrer la même moitié sur chaque demi-écran à chaque présentation.

Comme avec la méthode DSCQS, on présentera chaque paire d'images deux fois de suite, la première fois pour permettre aux visionneurs de l'observer et de l'examiner soigneusement, la seconde pour la confirmation et la notation. Chaque séquence durera entre 15 et 30 s. Les séquences pourront être précédées d'une «étiquette», ce qui facilitera le travail des visionneurs (voir la Fig. 3-6, représentant une paire d'images présentées sans effet miroir).

FIGURE 3-6

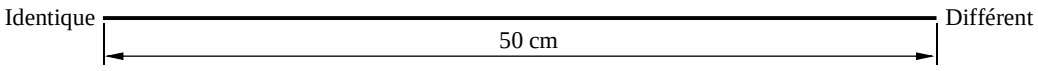
Paire d'images présentées sans effet miroir



A5-7.1.1.2 Échelle d'évaluation

Le critère d'acceptation d'une application LSDI est que la séquence d'essai (image comprimée) ne puisse pas être distinguée de la référence. L'évaluation des systèmes à l'essai peut se faire au moyen des nombreuses méthodes d'évaluation traditionnellement utilisées. On propose par exemple les échelles de comparaison de stimulus recommandées (Annexe 4 de la Partie 2). Comme exemple spécifique d'échelle d'évaluation, on propose l'échelle continue (sans catégorie) IDENTIQUE/DIFFÉRENT décrite au § A4-4.2 de l'Annexe 4 de la Partie 2.

FIGURE 3-7



BT.0500-03-7

A5-7.1.1.3 Session d'évaluation

La session, qui pourra comprendre plusieurs séances selon le nombre de conditions d'essai, portera sur deux types d'observations: observations de test et observations de vérification. Dans une observation de test, une moitié de l'écran affiche la référence et l'autre la séquence d'essai. Dans une observation de vérification, les deux moitiés de l'écran affichent la référence. La vérification permet de déterminer si le jugement de l'observateur est biaisé.

Pour chaque système à l'essai, les observations suivantes devront être faites sur chaque séquence:

TABLEAU 3-6

Partie gauche de l'affichage	Partie droite de l'affichage
Moitié gauche de la référence	Moitié gauche de la séquence d'essai
Moitié droite de la référence	Moitié droite de la séquence d'essai
Moitié gauche de la séquence d'essai	Moitié gauche de la référence
Moitié droite de la séquence d'essai	Moitié droite de la référence

De préférence, chaque configuration sera présentée au moins deux fois. Pour chaque système, et pour chaque séquence d'essai, les observations de vérification suivantes devront être faites:

TABLEAU 3-7

Panneau gauche	Panneau droit
Moitié gauche de la référence	Moitié gauche de la référence
Moitié droite de la référence	Moitié droite de la référence

Ici encore, chaque configuration sera présentée au moins deux fois de préférence.

Les séances de la session d'essai ne devront pas durer plus d'une heure, et une pause de 15 min sera prévue entre deux séances consécutives. Les observations d'essai et les observations de vérification faites sur les combinaisons sortie codec/séquence d'essai seront réparties entre les séances de façon pseudo-aléatoire. Malgré le surcroît de complexité qui en résultera, il sera préférable de prévoir certaines restrictions dans la procédure. Par exemple, dans une session comportant quatre séances, on pourrait affecter de façon aléatoire chacune des quatre observations d'essai correspondant à une combinaison signal codec/signal d'essai à une position choisie de façon aléatoire dans l'une des séances. Cette méthode présente l'avantage de garantir une bonne répartition, sur la totalité de la session, de toutes les observations d'essai.

A5-7.1.1.4 Traitement des notes d'évaluation

Pour une observation donnée, la note est la distance entre la position «IDENTIQUE» de l'échelle et la marque faite par l'observateur sur une échelle de 0-100. Les résultats sont convertis en notes moyennes d'opinion qui permettent de classer les systèmes testés. Selon le nombre d'observations par système (observateurs × séquences d'essai × répétitions), on pourra procéder à une analyse de variance (ANOVA)[7]. Les résultats des observations de vérification pourront servir à déterminer un «seuil de hasard».

A5-7.1.2 Phase II

La Phase II a notamment pour objet de donner une meilleure précision dans le classement relatif des résultats de la Phase I, dont la précision et la fiabilité dépendent du nombre d'observateurs et/ou des séances d'évaluation. Autre objectif important, il s'agit de susciter des commentaires en ce qui concerne les éléments sur lesquels les images sont perçues comme différentes et sur lesquels les évaluations de la Phase I ont été faites.

Dans la Phase II, les experts examinent les séquences qui ont été visionnées et, dans ce cas, ils peuvent formuler des commentaires sur les séquences projetées, visionner telle ou telle partie d'une séquence ou la totalité d'une séquence autant de fois qu'ils le souhaitent, soit pour l'examiner soit pour démontrer un point, ce qui leur permet de parvenir à un consensus dans leur évaluation et de bien décrire ce qu'ils voient. Le recours aux fonctions «ralenti», «décomposition du mouvement» et «arrêt sur image» est même autorisé si le visionneur le souhaite. Ici, les opérations appelleront parfois l'intervention du responsable de l'essai, qui pourra les coordonner.

A5-7.1.2.1 Regroupement des séquences d'essai

Pour le bon déroulement de la Phase II, il faudra regrouper les séquences d'essai par contenu en différents jeux de séquences de base pour visionnage d'experts (BES, *basic expert viewing set*) comportant les séquences codées dérivées d'une même source et ordonnées en fonction du classement établi pendant la Phase I.

Les séquences d'essai seront classées dans l'ordre croissant des notes moyennes d'opinion. Il y aura autant d'ensembles BES que de séquences différentes utilisées pendant l'essai.

A5-7.1.2.2 Sous-groupe de visionnage d'experts des séquences de base

Pendant une séance de sous-groupe de visionnage d'experts des séquences de base (BEV, *basic expert viewing*), les experts examinent toutes les séquences d'un ensemble de visionnage de base pour experts donné pour confirmer ou modifier le classement établi à l'issue des essais formels de la Phase I: c'est donc l'occasion de confirmer ou de modifier la visibilité relative des différences.

A5-7.1.2.3 Plan de Phase II

Pendant la Phase II, tous les visionnages d'experts des séquences de base (BEV) doivent être effectués. Les experts seront informés de ce que l'ordre de présentation correspond à l'ordre de classement établi pendant la Phase I, mais ils n'auront aucun élément d'information concernant le classement des systèmes proposés.

La Phase II sera un travail de groupe à l'issue duquel les experts formuleront des avis consensuels.

Avant le début de la Phase II, les visionneurs recevront, si possible par écrit, des instructions leur demandant:

---

[7] Un total de 10 à 20 observations concernant un élément de moindre importance suffit à justifier un traitement statistique déductif, une analyse de variance par exemple.

– d'examiner toutes les séquences couvertes par le BEV considéré;
– de commenter l'ordre de classement des séquences couvertes par le BEV considéré; en cas de désaccord, un nouveau classement sera établi;
– de commenter chaque cas en formulant des observations détaillées sur la nature des différences éventuellement observées;
– de documenter leur classement, leurs commentaires et leurs observations.

Il appartiendra au responsable de l'essai de rassembler toutes les observations formulées par les groupes et de repérer les divergences éventuelles. Pendant les tests proprement dits, les résultats obtenus par les différents groupes à l'issue de la Phase I et de la Phase II resteront confidentiels, de telle sorte que les groupes suivants ne puissent pas être influencés.

Lorsque cela sera possible, le responsable de l'essai pourra signaler les divergences et rechercher une solution au problème en demandant d'autres évaluations sur les points controversés. Cette dernière disposition a pour objet d'assurer un consensus général.

A5-8 Rapport

Le rapport final de la série d'essais, établi par le responsable, comprendra les informations suivantes:

– résultats de la Phase I (y compris les tableaux de notes moyennes d'opinion, et résultats, selon le cas, des analyses statistiques effectuées);
– observations formulées par les experts pendant la Phase II;
– observations relatives à toute réévaluation d'un classement;
– toutes informations utiles sur les conditions d'observation, les caractéristiques des signaux source, le traitement du signal, les caractéristiques du projecteur, les réglages du projecteur, les données de réglage chromatique, la sélection des visionneurs et les conditions d'essais proprement dites;
– description détaillée des caractéristiques du système d'affichage (courbes de moyenne des temps entre défaillances, etc.);
– résumé et conclusions.

Annexe 6
de la Partie 3

Méthode d'évaluation subjective de la qualité vidéo dans les applications multimédias

A6-1 Introduction

De nombreux pays ont commencé à mettre en place des systèmes de radiodiffusion numérique permettant d'offrir des applications de radiodiffusion multimédias et de données comportant des signaux vidéo, des signaux audio, des images fixes, du texte et des éléments graphiques.

Des méthodes normalisées d'évaluation subjective sont nécessaires pour spécifier la qualité de fonctionnement requise et pour vérifier que les solutions techniques envisagées pour chaque application conviennent. Des méthodes subjectives sont nécessaires car elles donnent des mesures qui permettent aux industriels d'anticiper plus directement les réactions des utilisateurs finals.

Le système de radiodiffusion nécessaire pour fournir des applications multimédias est sensiblement différent de celui qui est actuellement utilisé: accès aux informations par le biais de récepteurs fixes et/ou mobiles; débit d'images pouvant être fixe ou variable; grande variété de tailles d'image possibles (à savoir, de SQCIF à HDTV); signaux audio, texte et/ou signaux sonores généralement imbriqués dans l'image vidéo; possibilité de traitement de l'image vidéo au moyen de codecs vidéo évolués; et forte dépendance de la distance d'observation préférée vis-à-vis de l'application.

Il convient d'appliquer dans ce nouveau contexte les méthodes d'évaluation subjective spécifiées dans la Partie 2. En outre, de nouvelles méthodes pourraient être envisagées pour les systèmes multimédias afin de répondre aux besoins des utilisateurs en ce qui concerne les caractéristiques du domaine multimédia.

La présente Annexe décrit l'évaluation subjective non interactive de la qualité vidéo dans les applications multimédias. Ces méthodes peuvent notamment être utilisées pour choisir des algorithmes, déterminer la catégorie d'un système audiovisuel en fonction de ses performances ou évaluer le niveau de qualité vidéo pendant une connexion audiovisuelle.

A6-2 Description générale

A6-2.1 Conditions d'observation

Les conditions d'observation recommandées sont données dans le Tableau 3-8. La taille et le type de l'affichage utilisé devraient être choisis en fonction de l'application considérée. Comme plusieurs technologies d'affichage doivent être utilisées dans les applications multimédias, il convient d'indiquer toutes les informations utiles concernant l'affichage (par exemple, fabricant, modèle et spécifications) utilisé pour l'évaluation.

Lorsque des systèmes utilisant des PC sont utilisés pour présenter les séquences, il convient d'indiquer aussi les caractéristiques de ces systèmes (par exemple, carte d'affichage vidéo).

Le Tableau 3-9 donne un exemple de paramètres de configuration du système multimédia à évaluer.

Si les images d'essai sont obtenues à partir d'un décodeur-lecteur combiné particulier, il faut ôter l'enveloppe propriétaire pour obtenir un affichage anonyme. Cette opération est nécessaire pour garantir que l'évaluation de la qualité n'est pas influencée par la connaissance de l'environnement d'origine.

Lorsque les systèmes évalués dans un essai utilisent un format d'image réduit (par exemple CIF, SIF ou QCIF, etc.), les séquences devraient être affichées dans une fenêtre de l'écran. L'arrière-plan de l'écran devrait être de couleur grise 50%.

TABLEAU 3-8

Conditions d'observation recommandées à utiliser pour l'évaluation de la qualité dans les applications multimédias

Paramètre	Valeur
Distance d'observation[1)]	Valeur imposée: 1-8 H Pas de valeur imposée: suivant la préférence de l'observateur
Luminance de crête de l'écran	70-250 cd/m^2
Rapport de la luminance de l'écran inactif à la luminance de crête	≤ 0,05
Rapport de la luminance de l'écran, quand on ne reproduit que le niveau du noir dans une salle complètement obscure, à celle qui correspond au blanc maximal	≤ 0,1
Rapport de la luminance de l'arrière-plan, derrière l'écran, à la luminance de crête de l'image[(2)]	≤ 0,2
Chromaticité de l'arrière-plan[(3)]	D_{65}
Éclairement d'ambiance de la salle[(2)]	≤ 20 lux

(1) La distance d'observation dépend généralement de l'application.

(2) La valeur indiquée permet une détectabilité maximale des distorsions; pour certaines applications, des valeurs supérieures sont autorisées ou sont déterminées par l'application.

(3) Pour les écrans de PC, la chromaticité de l'arrière-plan devrait être la plus proche possible de la chromaticité du «point blanc» de l'affichage.

TABLEAU 3-9

Configuration du système multimédia à évaluer

Paramètre	Spécification
Type d'affichage	
Taille de l'affichage	
Carte d'affichage vidéo	
Fabricant	
Modèle	
Informations concernant l'image	

A6-2.2 Signaux source

Le signal source fournit directement l'image de référence et les signaux d'entrée pour le système à évaluer. La qualité des séquences source devrait être aussi élevée que possible. Le signal vidéo devrait en principe être enregistré dans des fichiers multimédias utilisant yuv (formats 4:2:2, 4:4:4) ou RGB (24 ou 32 bits). Lorsque l'expérimentateur souhaite comparer les résultats obtenus par différents laboratoires, il est nécessaire d'utiliser un ensemble commun de séquences source afin d'éliminer une autre source de variation.

A6-2.3 Choix du matériel d'essai

Le nombre et le type des scènes d'essai sont très importants pour l'interprétation des résultats de l'évaluation subjective. Pour certains processus, les dégradations observées sur la plupart des séquences peuvent être plus ou moins identiques. Dans ces conditions, les résultats obtenus à partir d'un petit nombre de séquences (par exemple, deux) devraient rester significatifs. Toutefois, l'impact des nouveaux systèmes dépend souvent pour beaucoup de la scène et du contenu de la séquence. En pareil cas, le nombre et le type des scènes d'essai devraient être choisis de manière à pouvoir procéder à une généralisation raisonnable pour les programmes normaux. En outre, il convient de choisir un matériel «critique mais sans excès» pour le système à évaluer. Par «sans excès» on entend que la scène pourra très bien faire partie de programmes de télévision normaux. Les caractéristiques spatiales et temporelles perçues pour une scène pourraient donner une indication utile de la complexité de cette scène. Des mesures des caractéristiques spatiales et temporelles perçues sont présentées plus en détail dans l'Annexe 6 de la Partie 1.

A6-2.4 Gamme de conditions et ancrage

Étant donné que la plupart des méthodes d'évaluation sont sensibles aux variations de la gamme et de la distribution des conditions observées, les séances d'évaluation subjective doivent inclure les gammes complètes de variation des facteurs. On peut atteindre toutefois plus ou moins le même objectif avec une gamme plus restreinte en présentant également certaines conditions qui se situeront aux extrémités des échelles. Elles peuvent être représentées comme exemples et identifiées comme étant les plus extrêmes (ancrage direct) ou réparties tout au long de la séance et non identifiées comme étant les plus extrêmes (ancrage indirect). Il convient, si possible, d'utiliser une large gamme de qualité.

A6-2.5 Observateurs

Il convient de sélectionner au moins 15 observateurs qui ne seront ni des spécialistes, en ce sens qu'ils ne s'occupent pas directement, dans le cadre de leur travail habituel, des questions liées à la qualité des images, ni des observateurs expérimentés. Avant chaque séance, les observateurs seront sélectionnés à l'aide de mires de Snellen ou de Landolt pour leur acuité visuelle normale ou rendue normale par correction et leur vision normale des couleurs, cela à l'aide de mires choisies à cet effet (d'Ishihara, par exemple).

Le nombre d'observateurs dépend de la sensibilité et de la fiabilité de la procédure d'essai retenue ainsi que de l'ampleur escomptée de l'effet évalué.

Les expérimentateurs devraient donner le plus de détails possibles sur les caractéristiques des groupes d'évaluation qu'ils ont retenus afin d'étudier plus avant ce facteur. Ils pourraient donner des précisions sur l'activité professionnelle (par exemple, fonctionnaire d'un organisme de radiodiffusion, étudiant d'une université, personnel de bureau), le sexe et l'âge.

A6-2.6 Modèle expérimental

Le soin est laissé à l'expérimentateur de choisir un modèle expérimental en fonction d'objectifs spécifiques en termes de coût et de précision. Il est préférable d'inclure au moins deux reproductions (c'est-à-dire répétitions de conditions identiques) dans l'expérience. Les reproductions permettent de calculer la fiabilité individuelle et, si nécessaire, d'ignorer des résultats non fiables issus de certains observateurs. Elles permettent aussi de compenser dans une certaine mesure les effets d'apprentissage dans le cadre d'un essai donné. On obtiendra une autre amélioration du traitement des effets d'apprentissage en prévoyant quelques «présentations fictives» au début de chaque séance d'essai. Il conviendra de choisir des conditions représentatives des présentations qui seront utilisées plus tard pendant la séance. Les présentations préliminaires ne sont pas à prendre en compte dans l'analyse statistique des résultats d'essai.

Une séance, c'est-à-dire une série de présentations, ne devrait pas durer plus d'une demi-heure.

Lorsque plusieurs scènes ou algorithmes sont évalués, ils devraient être présentés dans un ordre aléatoire. Cet ordre aléatoire pourrait être modifié pour faire en sorte que les mêmes scènes ou les mêmes algorithmes ne soient pas présentés à des moments proches (c'est-à-dire consécutivement).

A6-3 Méthodes d'évaluation

La qualité vidéo des systèmes multimédias peut être examinée au moyen des méthodes décrites dans la Partie 2.

La méthode d'évaluation subjective de la qualité vidéo multimédia (SAMVIQ) tire parti des caractéristiques du domaine multimédia et peut être utilisée pour évaluer les performances des systèmes multimédias.

Annexe 7
de la Partie 3

Évaluation subjective des systèmes de TV3D stéréoscopique

A7-1 Dimensions (de perception) d'évaluation

La TV3D stéréoscopique exploite les caractéristiques du système visuel binoculaire humain en recréant les conditions permettant de percevoir la profondeur relative des objets dans la scène visuelle. Actuellement, l'imagerie stéréoscopique nécessite principalement de prendre au moins deux vues de la même scène depuis deux caméras alignées horizontalement. Les images des objets de la scène auront des positions relatives différentes dans la vue de gauche et dans celle de droite. Cette différence de positions relatives dans les deux vues, généralement appelée disparité d'image (ou parallaxe), est habituellement exprimée par un nombre de pixels, une distance physique (par exemple en mm) ou une mesure relative (par exemple en pourcentage de la largeur d'écran). Il convient de distinguer la disparité d'image de la disparité angulaire (rétinienne). De fait, les mêmes informations de disparité d'image produisent des disparités angulaires (rétiniennes) différentes pour des distances d'observation différentes. L'amplitude et le sens de la profondeur perçue dépendent de l'amplitude et du sens des disparités rétiniennes provoquées par l'image stéréoscopique.

Les facteurs généralement utilisés pour l'évaluation des images de télévision monoscopiques, tels que la résolution, la restitution des couleurs, la restitution du mouvement, la qualité générale, la netteté, etc. pourraient aussi être utilisés pour les systèmes de télévision stéréoscopique. À cela s'ajouteraient de nombreux facteurs propres aux systèmes de télévision stéréoscopique, par exemple la résolution en profondeur, qui est la résolution spatiale dans le sens de la profondeur, le mouvement en profondeur, facteur qui indique si le mouvement dans le sens de la profondeur est restitué sans à-coups, et les distorsions spatiales, dont deux exemples sont bien connus: l'*effet spectacle de marionnettes* – lorsque des objets sont perçus comme étant anormalement grands ou petits – et l'*effet carton* – lorsque des objets sont perçus stéréoscopiquement mais paraissent anormalement minces.

Il est possible d'identifier trois dimensions de perception principales qui, ensemble, ont une incidence sur la qualité offerte par un système stéréoscopique: la *qualité de l'image*, la *qualité de la profondeur* et le *confort visuel*. D'après certains chercheurs, on pourrait aussi mesurer l'incidence psychologique des technologies d'imagerie stéréoscopique en utilisant des concepts plus généraux tels que le *naturel* et l'*impression de présence.*

A7-1.1 Dimensions de perception principales

La *qualité de l'image* désigne la qualité perçue de l'image fournie par le système. Principale composante de la qualité de fonctionnement d'un système vidéo, elle dépend essentiellement des paramètres techniques et des erreurs introduites par, entre autres, les processus de codage et/ou de transmission.

La *qualité de la profondeur* désigne la capacité du système à renforcer l'impression de profondeur. La présence d'indices monoculaires (perspective linéaire, flou, gradients, etc.) permet de percevoir une certaine profondeur même dans des images 2D standards. Toutefois, les images 3D stéréoscopiques contiennent aussi des informations de disparité qui offrent des informations supplémentaires de profondeur et, par là-même, renforcent l'impression de profondeur par rapport aux images 2D.

Le *confort visuel* (l'*inconfort visuel*) désigne la sensation subjective de confort (d'inconfort) associée à l'observation d'images stéréoscopiques. Des images stéréoscopiques mal prises ou mal affichées peuvent créer beaucoup d'inconfort.

A7-1.2 Autres dimensions de perception

Le *naturel* désigne la perception de l'image stéréoscopique comme étant une représentation fidèle de la réalité (perception de réalisme). L'image stéréoscopique peut présenter différents types de distorsions qui la rendent moins naturelle. Par exemple, certains objets stéréoscopiques sont parfois perçus comme étant anormalement grands ou petits (effet spectacle de marionnettes), ou ils paraissent anormalement minces (effet carton).

L'*impression de présence* désigne l'impression subjective d'être à un endroit différent de l'endroit où l'on se trouve.

La présente Recommandation donne des informations sur les méthodes et les procédures d'évaluation des trois dimensions principales – qualité de l'image, qualité de la profondeur et confort visuel – décrites ci-dessus. Les méthodes d'évaluation du naturel et de l'impression de présence n'y sont pas traitées, mais il est prévu de les aborder dans une future version.

A7-2 Méthodes subjectives

La présente Recommandation décrit de nombreuses méthodes d'évaluation de la qualité d'image. Dans toutes les méthodes, un ensemble de séquences vidéo, traitées par les systèmes (par exemple un algorithme avec différents paramètres, une technologie de codage à différents débits binaires, différents scénarios de transmission, etc.) à l'étude, est présenté à un groupe d'observateurs dans une série d'expériences d'évaluation. Dans chaque expérience, il est demandé aux observateurs d'évaluer une caractéristique particulière (par exemple la qualité de l'image) de la ou des séquences vidéo sur une échelle donnée. Les principales différences entre les diverses méthodes résident dans le mode de présentation, autrement dit la manière dont les séquences vidéo sont présentées aux observateurs, et dans l'échelle utilisée par les observateurs pour noter ces séquences.

Les images de test sont des images stéréo binoculaires choisies sur la base des éléments décrits au § A7-4. Les observateurs évaluent les trois éléments suivants:

– la qualité de l'image: l'effet produit sur la résolution d'images 3D stéréoscopiques par un système opérant entre les images de test et l'écran utilisé pour afficher les images à évaluer;
– la qualité de la profondeur: l'effet produit sur la perception de la profondeur d'images 3D stéréoscopiques par un système opérant entre les images de test et l'écran utilisé pour afficher les images à évaluer;
– le confort visuel: l'effet produit sur la facilité d'observation d'images 3D stéréoscopiques par un système opérant entre les images de test et l'écran utilisé pour afficher les images à évaluer.

La présente Annexe reprend six méthodes de la présente Recommandation; ces méthodes ont donné de bons résultats au cours des vingt dernières années dans des travaux de recherche concernant la qualité de l'image, la qualité de la profondeur et le confort visuel en imagerie stéréoscopique. Il s'agit de:

– la méthode à un seul stimulus (SS);
– la méthode à double stimulus utilisant une échelle de dégradation (DSIS);
– la méthode à double stimulus utilisant une échelle de qualité continue (DSCQS);
– la méthode de comparaison de stimulus (SC);
– la méthode d'évaluation continue de la qualité avec stimulus unique (SSCQE);
– la méthode d'évaluation continue à double stimulus simultanés (SDSCE).

Les méthodes ont parfois été utilisées moyennant de légères modifications, par exemple des échelles différentes ont été utilisées pour le confort visuel. Le mode de présentation et les échelles associés à la méthode d'évaluation de la qualité de l'image, de la qualité de la profondeur et du confort visuel sont récapitulés respectivement dans les Tableaux 3-10, 3-11 et 3-12.

On trouvera ci-après une brève description de chaque méthode, puis une description des éléments communs à toutes les méthodes.

TABLEAU 3-10

Méthode d'évaluation subjective de la qualité de l'image

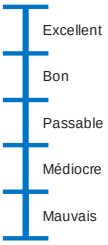
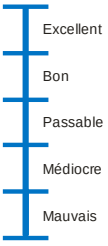
Mode de présentation	Durée de la séquence	Échelle binaire	Échelle discrète	Échelle continue
Méthode à un seul stimulus (SS), telle que décrite au § 6.1 de l'Annexe 1.	~10 s		5 Excellent 4 Bon 3 Passable 2 Médiocre 1 Mauvais	
Méthode à double stimulus utilisant une échelle de dégradation (DSIS); telle que décrite au § 4 de l'Annexe 1.			5 Imperceptible 4 Perceptible, mais non gênant 3 Légèrement gênant 2 Gênant 1 Très gênant	
Méthode à double stimulus utilisant une échelle de qualité continue (DSCQS), telle que décrite au § 5 de l'Annexe 1.	~10 s			

TABLEAU 3-10 (*fin*)

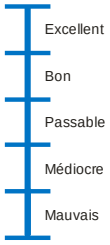
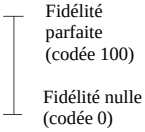
Mode de présentation	Durée de la séquence	Échelle binaire	Échelle discrète	Échelle continue
Méthode de comparaison de stimulus (SC), telle que décrite au § 6.2 de l'Annexe 1.	~10 s	A en fonction de B	−3 Beaucoup moins bon −2 Moins bon −1 Un peu moins bon 0 Équivalent 1 Légèrement meilleur 2 Meilleur 3 Bien meilleur	
Méthode d'évaluation continue de la qualité avec stimulus unique (SSCQE), telle que décrite au § 6.3 de l'Annexe 1.	~3-5 min			
Méthode d'évaluation continue à double stimulus simultanés (SDSCE), telle que décrite au § 6.4 de l'Annexe 1.				

TABLEAU 3-11

Méthode d'évaluation subjective de la qualité de la profondeur

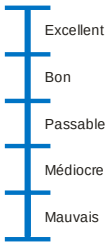
Mode de présentation	Durée de la séquence	Échelle binaire	Échelle discrète	Échelle continue
Méthode à un seul stimulus (SS), telle que décrite au § 6.1 de l'Annexe 1.	~10 s		5 Excellent 4 Bon 3 Passable 2 Médiocre 1 Mauvais	
Méthode à double stimulus utilisant une échelle de dégradation (DSIS), telle que décrite au § 4 de l'Annexe 1			5 Imperceptible 4 Perceptible, mais non gênant 3 Légèrement gênant 2 Gênant 1 Très gênant	

TABLEAU 3-11 (*fin*)

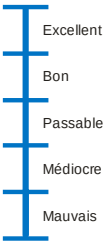
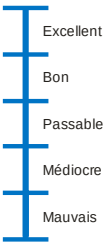
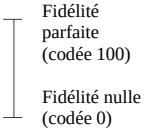
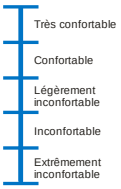
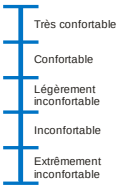
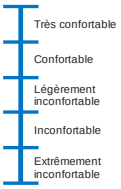
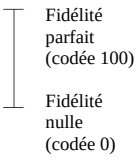
Mode de présentation	Durée de la séquence	Échelle binaire	Échelle discrète	Échelle continue
Méthode à double stimulus utilisant une échelle de qualité continue (DSCQS), telle que décrite au § 5 de l'Annexe 1.	~10 s			
Méthode de comparaison de stimulus (SC), telle que décrite au § 6.2 de l'Annexe 1.	~10 s	A en fonction de B	−3 Beaucoup moins bon −2 Moins bon −1 Un peu moins bon 0 Équivalent 1 Légèrement meilleur 2 Meilleur 3 Bien meilleur	
Méthode d'évaluation continue de la qualité avec stimulus unique (SSCQE), telle que décrite au § 6.3 de l'Annexe 1.	~3-5 min			
Méthode d'évaluation continue à double stimulus simultanés (SDSCE), telle que décrite au § 6.4 de l'Annexe 1.				

TABLEAU 3-12

Méthode d'évaluation subjective du confort visuel

Mode de présentation	Durée de la séquence	Échelle binaire	Échelle discrète	Échelle continue
Méthode à un seul stimulus (SS), telle que décrite au § 6.1 de l'Annexe 1.	~10 s		5 Très confortable 4 Confortable 3 Légèrement inconfortable 2 Inconfortable 1 Extrêmement inconfortable	
Méthode à double stimulus utilisant une échelle de dégradation (DSIS), telle que décrite au § 4 de l'Annexe 1			5 Imperceptible 4 Perceptible, mais non gênant 3 Légèrement gênant 2 Gênant 1 Très gênant	
Méthode à double stimulus utilisant une échelle de qualité continue (DSCQS), telle que décrite au § 5 de l'Annexe 1.	~10 s			
Méthode de comparaison de stimulus (SC), telle que décrite au § 6.2 de l'Annexe 1.	~10 s	A en fonction de B	−3 Beaucoup moins bon −2 Moins bon −1 Un peu moins bon 0 Équivalent 1 Légèrement meilleur 2 Meilleur 3 Bien meilleur	
Méthode d'évaluation continue de la qualité avec stimulus unique (SSCQE), telle que décrite au § 6.3 de l'Annexe 1.	~3-5 min			
Méthode d'évaluation continue à double stimulus simultanés (SDSCE), telle que décrite au § 6.4 de l'Annexe 1.				

A7-3 Conditions générales d'observation

Les conditions d'observation (luminance et contraste de l'écran, éclairement du fond, distance d'observation, etc.) devraient être cohérentes avec celles utilisées pour la 2D qui sont décrites au § 2.1 de la Partie 1. Deux arguments militent en faveur de cette cohérence. Le premier est que, dans la pratique, les utilisateurs regarderont la TV3D sur les mêmes écrans et dans les mêmes conditions d'observation que pour la 2D. Le deuxième est que les progrès des technologies vidéo de TV3D devront souvent être mesurés par rapport aux progrès des technologies vidéo de la TVHD standard.

Le paragraphe 2.1 de la Partie 1 spécifie deux critères possibles pour le choix de la distance d'observation. Il convient de choisir la distance d'observation nominale. Pour un système numérique, la distance d'observation nominale est la distance à laquelle deux pixels adjacents forment un angle de 1 minute d'arc depuis l'œil de l'observateur.

Il convient de noter que, étant donné que deux pixels adjacents forment un angle de 1 minute d'arc depuis l'œil de l'observateur, la plus petite disparité angulaire (rétinienne) qui peut être représentée par le système (c'est-à-dire la résolution en profondeur du système) à la distance d'observation nominale est égale à 1 minute d'arc (ou, de façon équivalente, à 60 secondes d'arc). D'après des travaux de recherche, près de 97% de la population peut distinguer des disparités horizontales égales ou inférieures à 140 secondes d'arc, et au moins 80% peut détecter des disparités horizontales de 30 secondes d'arc. Par conséquent, la plupart des observateurs devraient pouvoir détecter la plus petite disparité représentable dans les systèmes vidéo 3D actuels à la distance d'observation nominale.

A7-4 Séquences de test

Le choix des séquences de test devrait dépendre de la question expérimentale traitée dans l'étude. D'une manière générale, le contenu des séquences de test (sport, théâtre, film, etc.) et leurs caractéristiques spatio-temporelles devraient être représentatifs des programmes diffusés par le service étudié.

En outre, le contenu des séquences de test stéréoscopiques choisies devrait aussi être relativement confortable à regarder. Le confort visuel des images stéréoscopiques dépend essentiellement des disparités d'image (parallaxe) et des conditions d'observation. Il convient donc de veiller à ce que les disparités ne dépassent pas les limites indiquées dans le paragraphe qui suit, sauf si l'étude vise expressément à mesurer le confort visuel. De plus, il convient, chaque fois que possible, de mesurer et d'indiquer la moyenne, l'écart type et la plage (min/max) de la distribution de disparité des séquences de test.

La parallaxe, les incohérences entres les images de gauche et de droite, et la distribution et la modification de parallaxe sont des éléments dont il convient de tenir compte lors du choix d'images 3D stéréoscopiques faciles à observer en vue d'une évaluation. La relation entre une image 3D stéréoscopique facile à observer et la parallaxe, les incohérences entres les images de gauche et de droite, et la distribution et la modification de parallaxe est décrite dans les paragraphes qui suivent.

A7-4.1 Utilisation des séquences video de référence

Si la séquence de référence est disponible, les chercheurs souhaiteront peut-être l'inclure dans l'ensemble des séquences de test. La référence est généralement une version de la séquence de test qui n'a subi aucun traitement (autrement dit la séquence source d'origine). Pour les études stéréoscopiques, la principale référence est la séquence stéréoscopique non traitée d'origine. Le plan d'expérience pourrait toutefois aussi inclure la version monoscopique de la référence (autrement dit une seule vue de la séquence source d'origine); par exemple, pour l'étude du confort visuel, il pourrait être utile de faire une comparaison avec la référence monoscopique. La version monoscopique de la référence devrait être présentée en mode 3D (par exemple la vue de gauche est présentée à la fois à

l'œil gauche et à l'œil droit et on utilise les mêmes paramètres matériels 3D que pour la vraie séquence stéréoscopique). L'inclusion de la référence dans le plan d'expérience comporte deux avantages importants: elle permet de mesurer la transparence (autrement dit la fidélité) offerte par l'algorithme ou la technologie à l'étude[8] et elle sert de repère de qualité élevée afin d'aider à stabiliser les notes[9].

A7-4.2 Limites de confort visuel

Une disparité/parallaxe excessive entraîne un inconfort visuel, peut-être en raison de l'accentuation du conflit entre accommodation et vergence. Il a donc été proposé, pour réduire ce conflit, de faire en sorte que les disparités dans l'image stéréoscopique soient suffisamment faibles pour que les profondeurs perçues des objets se situent dans les limites d'une «zone de confort». Pour définir ces limites, plusieurs approches ont été proposées. L'une des approches consiste à utiliser une mesure de la parallaxe relative à l'écran, exprimée en pourcentage de la largeur de l'écran. Des valeurs de 1% pour les disparités croisées/négatives et de 2% pour les disparités non croisées/positives (pour une valeur totale d'environ 3%) ont été proposées. Selon une autre approche, on utilise la profondeur de champ de l'œil pour délimiter la zone de confort. Pour les conditions d'observation types de programmes de télévision, les chercheurs ont pris comme hypothèse une profondeur de champ comprise entre ±0,2D (dioptries) et ±0,3D (dioptries). Pour une résolution d'image de TVHD de 1 920 × 1 080 (Recommandation UIT-R BT.709) et une distance d'observation nominale de 3,1H, ces valeurs correspondent approximativement à une parallaxe relative à l'écran de ±2% et de ±3%. Enfin, une troisième approche consiste à spécifier les limites de confort en fonction de la disparité rétinienne et à fixer ces limites à ±1° d'angle visuel à la fois pour les disparités positives et pour les disparités négatives.

Il est à noter que ces différentes approches ont tendance à converger vers les mêmes limites de confort. Rappelons que, à la distance d'observation nominale, deux pixels adjacents forment un angle de 1 minute d'arc depuis l'œil de l'observateur. Donc 60 pixels correspondent à 1° d'angle visuel. On peut ainsi spécifier facilement les limites de confort en termes de disparité rétinienne (pour un observateur moyen). Par exemple, pour une résolution d'image de TVHD de 1 920 × 1 080 (Recommandation UIT-R BT.709), 1% (~19,2 pixels) correspond approximativement à 20 minutes d'arc, 2% à ~40 minutes d'arc et 3% à ~60 minutes d'arc (ou, de façon équivalente, à 1°).

Il convient de noter que même si, à la distance d'observation nominale, deux pixels adjacents forment toujours un angle de 1 minute d'arc, la distance physique (par exemple en mm) entre ces pixels est plus grande pour des écrans plus grands (le nombre de pixels reste identique, mais les dimensions physiques de l'écran augmentent). Par conséquent, pour les limites plus élevées (par exemple ±3%) et pour les écrans plus grands, la distance physique entre les points correspondants (autrement dit la parallaxe des deux vues en mm) pourrait dépasser l'écartement des yeux de l'observateur moyen (~63-65 mm), ce qui pourrait donner lieu à un inconfort plus important.

---

[8] La transparence (fidélité) décrit le fait qu'un codec ou un système n'entraîne aucune dégradation par rapport à un système de transmission idéal. Il va de soi qu'on peut mesurer la transparence en comparant les notes attribuées à la séquence de référence et celles attribuées à la séquence traitée par l'algorithme ou la technologie à l'étude.

[9] Il est reconnu que la stabilité des notes dans l'espace (c'est-à-dire dans des laboratoires différents) et dans le temps (c'est-à-dire dans le même laboratoire mais à des moments différents) pourrait aussi être améliorée avec l'utilisation de repères de faible qualité. Cela étant, l'UIT a l'intention de produire/définir dès maintenant des repères normalisés de faible qualité pour l'évaluation des technologies d'imagerie stéréoscopique.

A7-4.3 Discordances entre les images de gauche et de droite

Dans les systèmes 3D stéréo, une image 3D binoculaire est formée par la présentation de l'image de gauche et de celle de droite à l'œil correspondant. Toute discordance entre ces deux images peut entraîner un stress psychophysique et, dans certains cas, l'observation en 3D peut échouer. Par exemple, lors de la prise de vues et de l'affichage de programmes de TV3D stéréoscopique, il peut y avoir des distorsions géométriques, comme une incohérence de taille, un décalage vertical ou une erreur de rotation, entre l'image de gauche et celle de droite. Il est préférable que les images de test ne comportent pas de telles distorsions géométriques. Voir le § 3.2.1 de l'Annexe 4 du Rapport UIT-R BT.2160-2 pour plus d'informations.

Les discordances entre l'image de gauche et celle de droite dont il convient de tenir compte lors du choix d'images 3D stéréoscopiques faciles à observer en vue d'une évaluation sont les suivantes:

– discordance géométrique (taille, déplacement vertical, rotation);

– discordance de brillance (niveau de blanc et de noir);

– diaphonie.

A7-4.4 Plage, distribution et modification de parallaxe

Dans le cas des images stéréoscopiques, les distributions de parallaxe sont corrélées avec le confort visuel.

La distribution de parallaxe d'images stéréoscopiques est discontinue pendant les trames de changement de scène. Les cas de parallaxe extrême ou de modification subite de parallaxe entraînent un inconfort visuel; il est donc important de prendre en considération avec soin la parallaxe des images de test. Voir le § 3.2.2 de l'Annexe 4 du Rapport UIT-R BT.2160-2 pour plus d'informations.

D'une manière générale, étant donné que les études utilisant des séquences de test stéréoscopiques peuvent donner lieu à un certain inconfort visuel, il est recommandé d'utiliser, chaque fois que possible, des séquences de test dont la disparité ne dépasse pas les limites de confort, même si des dépassements occasionnels peuvent être autorisés.

A7-5 Appareils expérimentaux

Les appareils expérimentaux (serveur vidéo, écran, etc.) devraient être en mesure d'afficher des séquences de test HD pleine résolution, par exemple en utilisant un format de mise en trame HDMI, ce qui permettrait d'élargir l'éventail des études qui peuvent être réalisées.

À ce jour, aucun écran de référence pour l'évaluation de la TV3D n'a été normalisé. On s'attend donc à ce que la plupart des chercheurs utilisent les écrans actuels de TV3D grand public. Étant donné que les caractéristiques de ces écrans peuvent varier d'un fabricant à l'autre, les chercheurs sont vivement encouragés à indiquer les paramètres de l'écran utilisé dans l'étude.

A7-6 Observateurs

A7-6.1 Taille de l'échantillon

Il est généralement recommandé de faire appel à au moins 30 observateurs. Il est toutefois reconnu que le nombre effectif dépendra des objectifs spécifiques de l'étude, tout en sachant que les considérations relatives à la taille de l'échantillon pour les études de la 3D ne sont pas différentes de celles utilisées pour les études de la 2D.

A7-6.2 Sélection des observateurs en fonction de leur vue

Il convient de sélectionner les observateurs en fonction de leur acuité visuelle, de leur perception des couleurs et de leur vision stéréoscopique au moyen de tests cliniques actuels de la vue, par exemple l'équivalent des mires de Snellen pour l'acuité visuelle, les planches d'Ishihara ou un équivalent pour les couleurs et le test de Randot ou un équivalent pour la vision stéréoscopique. Il est à noter que les tests de vision stéréoscopique tels que les tests de Randot, Stereo Fly ou Frisby mesurent généralement des disparités rétiniennes allant d'environ 20 à 400 secondes d'arc. Les chercheurs sont encouragés à indiquer les statistiques pertinentes concernant l'acuité stéréoscopique des observateurs participant à l'étude. Si une analyse plus détaillée de l'acuité stéréoscopique des participants est nécessaire, les chercheurs peuvent recourir aux images présentées dans la Pièce jointe 1 à la présente Annexe.

A7-7 Instructions à l'intention des observateurs

Les instructions devraient être adaptées aux dimensions (par exemple qualité de la profondeur, confort, etc.) à étudier. En particulier, les directives déontologiques pour les études de la 3D sont plus strictes que celles qui sont généralement utilisées pour l'évaluation de la qualité d'images 2D car les participants peuvent ressentir un inconfort visuel. D'une manière générale, pour les études portant sur la 3D, il faut veiller à bien informer les participants des motivations de chaque étude ainsi que des éventuelles conséquences négatives de l'exposition aux stimuli utilisés dans l'étude.

A7-8 Durée d'une séance

Si les séquences à observer sont censées être confortables, la durée de la séance de test peut être aussi longue que celle utilisée pour les études de la 2D (à savoir ~20-40 minutes entrecoupées de pauses). Si on sait que les séquences contiennent une parallaxe excessive et qu'elles sont donc susceptibles d'être inconfortables, la durée devrait être limitée.

A7-9 Variabilité des réponses

En général, les notes fournies par les observateurs dans les expériences d'évaluation subjective sont assez variables. Les différences entre les observateurs peuvent simplement être liées aux caractéristiques de la population de référence et, pour les atténuer, on peut augmenter la taille de l'échantillon.

Toutefois, une partie de la variabilité peut être liée à l'évolution des réponses fournies par les différents observateurs au cours de l'expérience. Cette évolution traduit une évolution des critères d'évaluation qui pourrait être due au fait d'acquérir davantage de pratique de la tâche, à un apprentissage des caractéristiques des artefacts, etc. Pour réduire autant que possible les effets négatifs de cette variabilité, les chercheurs devraient prévoir des procédures de formation appropriées (tâche, niveau de dégradation, etc.), recourir à de multiples randomisations (à savoir présenter les séquences de test dans différents ordres aléatoires aux différents observateurs), et répéter les présentations (ce qui permettrait aussi de mesurer l'éventuelle évolution des réponses fournies).

A7-10 Critères de rejet d'observateurs

Les critères de rejet d'observateurs (sélection des observateurs) pour les méthodes présentées au § A7-2 sont décrits dans la Partie 1.

A7-11 Analyse statistique

Les analyses statistiques pour l'étude des systèmes d'imagerie 3D sont les mêmes que pour les systèmes d'imagerie 2D.

Pièce jointe 1
à l'Annexe 7

Images pour l'examen de la vue

A7-1 Examen de la vue

Le Tableau 3-13 donne la liste des mires utilisées pour l'examen de la vue. Les 12 tests ont été choisis selon la hiérarchie du système visuel humain, depuis le plus bas niveau jusqu'au niveau le plus élevé. On trouvera ci-après la description de huit tests de la vue (VT, *vision test*) principaux, puis de quatre autres réservés à l'usage clinique. Les observateurs doivent avoir une stéréopsie normale, c'est-à-dire qu'ils doivent satisfaire au test VT-04 pour la stéréopsie fine et au test VT-07 pour la stéréopsie dynamique. Les six autres tests sont destinés à une caractérisation plus détaillée. Les mires doivent être regardées à une distance égale au triple de la hauteur de l'écran.

TABLEAU 3-13

Images stéréoscopiques pour l'examen de la vue

N°	Propriété testée	Objet du test	Contenu
1	Perception simultanée	Aptitude à percevoir simultanément, et dans la position correcte, des images en présentation dichoptique	L'image d'une cage est présentée à un œil et l'image d'un lion à l'autre œil
2	Fusion binoculaire	Aptitude à percevoir deux images dichoptiques dans l'œil gauche et l'œil droit sous la forme d'une image unique	L'image destinée à l'un des yeux comporte deux points, celle destinée à l'autre œil en comporte trois, avec un point en commun
3	Stéréopsie grossière	Aptitude à percevoir des images en présentation dichoptique avec une parallaxe, sous la forme d'une image unique donnant une impression de relief (profondeur) grossière	Les images destinées aux deux yeux sont une paire stéréo d'images représentant une libellule dont les ailes sont déployées
4	Stéréopsie fine	Aptitude à percevoir des images en présentation dichoptique avec une parallaxe, sous la forme d'une image unique donnant une impression de relief (profondeur) fine	On présente neuf losanges contenant chacun quatre cercles dont un seul a une petite parallaxe
5	Limite de fusion avec croisement	Aptitude à percevoir comme une image unique des images en présentation dichoptique avec des disparités croisées	On présente une paire stéréo de barres dont la parallaxe croisée varie de 10'/s
6	Limite de fusion sans croisement	Aptitude à percevoir comme une image unique des images en présentation dichoptique avec des disparités non croisées	On présente une paire stéréo de barres dont la parallaxe non croisée varie de 11'/s
7	Stéréopsie dynamique	Aptitude à percevoir la profondeur dans des images formées par des stéréogrammes en points aléatoires en mouvement	Stéréogramme dynamique en points aléatoires

TABLEAU 3-13 (*fin*)

N°	Propriété testée	Objet du test	Contenu
8	Acuité binoculaire	Acuité binoculaire, y compris la dissymétrie éventuelle de l'acuité monoculaire qui pourrait empêcher une bonne stéréopsie	Caractères E d'orientation et de dimensions variées
9	Strabisme horizontal	Déviation horizontale de l'œil que le patient ne parvient pas à maîtriser	Segment vertical et segment horizontal
10	Strabisme vertical	Déviation verticale de l'œil que le patient ne parvient pas à maîtriser	Segment vertical et segment horizontal
11	Aniséiconie	Perception d'images différentes en termes de taille et de forme par les deux yeux fixant un même objet	L'image de gauche est constituée des caractères «[o» et celle de droite des caractères «o]», les caractères «o» ayant la même position
12	Cyclophorie	Rotation d'un œil, ou de l'autre, autour de l'axe antéropostérieur lorsque la fusion est empêchée	L'image de gauche représente un cadran d'horloge et celle de droite, les aiguilles d'une horloge marquant 6 h

NOTE 1 – Ces images de test sont au format 1125/60/I (voir la Recommandation UIT-R BT.709).

NOTE 2 – On peut se procurer ces images de test auprès de l'Institute of Image Information and Television Engineers (ITE), 3-5-8 Shibakoen, Minato-ku, Tokyo 105-0011, Japon, téléphone: 81-3-3432-4675, courriel: ite@ite.or.jp.

On trouvera ci-après, côte à côte, les images miniatures de droite et de gauche destinées à la fusion libre avec croisement, aux fins d'explication.

1) *VT-01*: Perception simultanée (test du lion)

Test de l'aptitude à percevoir simultanément, et dans la position correcte, des images en présentation dichoptique. À un œil, on présente l'image d'une cage et à l'autre œil, l'image d'un lion dont la position se déplace de 12'/s. La taille de chaque image est fixée à 10°, ce qui permet aux observateurs de capter les images dans leur zone paramaculaire. Les observateurs dotés d'une vue normale peuvent voir le lion dans la cage à un certain instant de la période de présentation.

FIGURE 3-8

Mire pour le test VT-01

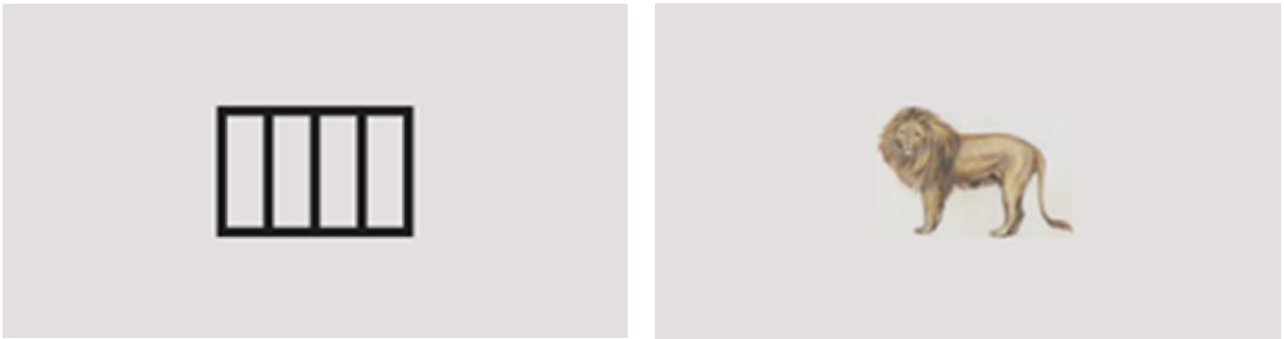


Image de droite Image de gauche

BT.0500-03-08

2) *VT-02*: Fusion binoculaire (test aux 4 points de Worth)

Test de l'aptitude à percevoir deux images dichoptiques dans l'œil gauche et l'œil droit comme une seule image. L'image destinée à l'un des yeux comporte deux points, celle destinée à l'autre œil en comporte trois, avec un point en commun. Les observateurs dotés d'une vue normale voient quatre points.

FIGURE 3-9

Mire pour le VT-02

Image de droite Image de gauche

BT.500-03-09

3) *VT-03*: Stéréopsie grossière (test de la libellule)

Test de l'aptitude à percevoir des images en présentation dichoptique avec une parallaxe, sous la forme d'une image unique donnant une impression de relief (profondeur) grossière. Les images destinées aux deux yeux sont une paire stéréo d'images représentant une libellule dont les ailes sont déployées. Les observateurs dotés d'une vue normale voient les ailes en avant de l'écran.

FIGURE 3-10

Mire pour le test VT-03

Image de droite Image de gauche

BT.0500-03-10

4) *VT-04*: Stéréopsie fine (test des cercles)

Test de l'aptitude à percevoir des images en présentation dichoptique avec une parallaxe, sous la forme d'une image unique donnant une impression de relief (profondeur) fine. On présente neuf losanges contenant chacun quatre cercles dont un seul a une petite parallaxe. Les observateurs dotés d'une vue normale voient le cercle avec une petite parallaxe en avant de l'écran. Le Tableau 3-14 donne le numéro des tests, les réponses correctes et l'angle de stéréopsie à la distance 3 *H*.

TABLEAU 3-14

Réponses correctes et parallaxe

Test N°	Réponses correctes	Angle de stéréopsie à la distance 3 *H* (")
1	En bas	480
2	À gauche	420
3	En bas	360
4	En haut	300
5	En haut	240
6	À gauche	180
7	À droite	120
8	À gauche	60
9	–	0

FIGURE 3-11

Mire pour le test VT-04

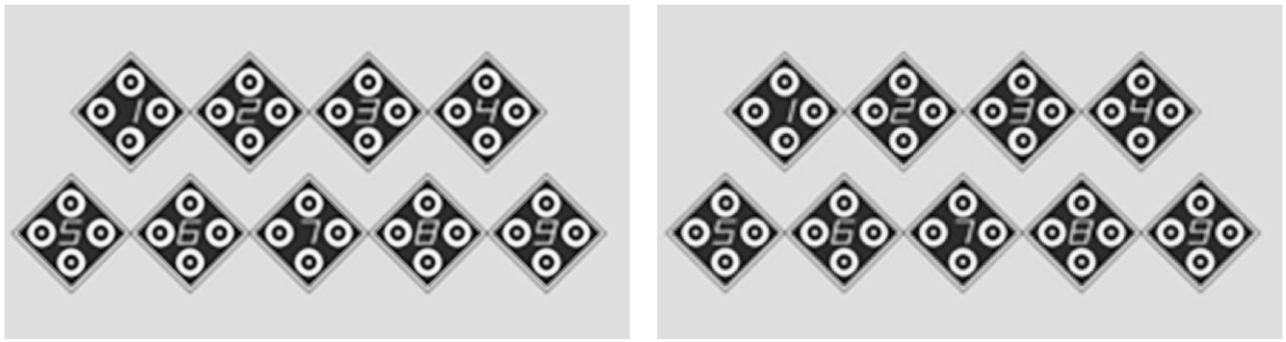


Image de droite Image de gauche

BT.0500-03-11

5) *VT-05*: Limite de fusion avec croisement (test des barres)

Test de l'aptitude à percevoir comme une image unique des images en présentation dichoptique avec des disparités croisées. On présente une paire stéréo de barres dont la parallaxe varie de 10'/s. Il est possible de mesurer les limites de la fusion pour les séries croissante et décroissante. Il est demandé aux observateurs de signaler leur rupture de fusion dès qu'ils voient deux images dans la série croissante et le rétablissement de la fusion dès qu'ils voient les images dichoptiques comme une image unique dans la série décroissante.

FIGURE 3-12

Mire pour le test VT-05

Image de droite Image de gauche

BT.0500-03-12

6) *VT-06*: Limite de fusion sans croisement (test des barres)

Test de l'aptitude à percevoir comme une image unique des images en présentation dichoptique avec des disparités non croisées. Les images présentées sont les mêmes que pour le test précédent avec croisement, mais les images de droite et de gauche sont interverties.

FIGURE 3-13

Mire pour le test VT-06

Image de droite Image de gauche

BT.0500-03-13

7) *VT-07*: Stéréopsie dynamique (test avec stéréogramme dynamique en points aléatoires)

Test de l'aptitude à percevoir la profondeur dans des images formées par des stéréogrammes en points aléatoires en mouvement. Les observateurs dotés d'une vue normale perçoivent une forme rectangulaire et un mouvement sinusoïdal en profondeur dans le stéréogramme.

FIGURE 3-14

Mire pour le test VT-07

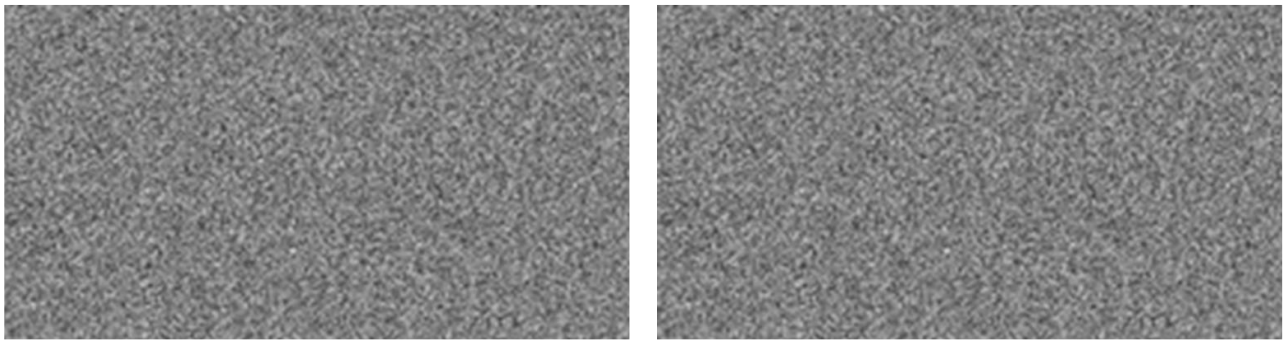


Image de droite Image de gauche

BT.0500-03-14

8) *VT-08*: Acuité binoculaire (test d'acuité)

Test de l'acuité binoculaire avec fusion binoculaire, y compris une dissymétrie éventuelle de l'acuité monoculaire qui pourrait empêcher une bonne stéréopsie. Les images contiennent quatre colonnes et cinq lignes composées de caractères E d'orientation et de dimensions variées. Les deux colonnes centrales peuvent être vues avec les deux yeux, les deux colonnes de gauche peuvent être vues seulement avec l'œil gauche, et les deux colonnes de droite seulement avec l'œil droit. Les observateurs dotés d'une vue normale peuvent distinguer correctement l'orientation des caractères E. Les dimensions des caractères correspondent à des acuités d'environ 1,0, 0,5, 0,33, 0,25 et 0,125 à la distance 3 H.

FIGURE 3-15

Mire pour le test VT-08

Image de droite Image de gauche

BT.0500-03-15

9 et 10) *VT-09*: Strabisme horizontal (test de Maddox horizontal) et *VT-10*: Strabisme vertical (test de Maddox vertical)

Ces mires permettent de mesurer la déviation horizontale et verticale de l'œil. La position des axes visuels l'un par rapport à l'autre est différente de celle requise par les conditions physiologiques. Les images sont constituées d'un segment vertical et d'un segment horizontal. Les observateurs dotés d'une vue normale voient le point d'intersection des segments aux alentours du milieu des segments. Le nombre de graduations donne la dioptrie prismatique pour un écartement des yeux de 65 mm à la distance 3,02 H.

FIGURE 3-16

Mire pour le test VT-09

BT.0500-03-16

FIGURE 3-17

Mire pour le test VT-10

BT.0500-03-17

11) *VT-11*: Aniséiconie (test des caractères «[ ]»)

Perception d'images différentes en termes de taille et de forme par les deux yeux fixant un même objet. L'image de gauche est constituée des caractères «[o» et celle de droite des caractères «o]», les caractères «o» ayant la même position. Les observateurs dotés d'une vue normale voient les caractères «[«et»]» comme ayant la même taille et la même hauteur.

FIGURE 3-18

Mire pour le test VT-11

BT.0500-03-18

12) *VT-12*: Cyclophorie (test de l'horloge)

Rotation d'un œil autour de l'axe antéropostérieur uniquement lorsqu'il est couvert et que la fusion est empêchée. L'image de gauche représente un cadran d'horloge et celle de droite, les aiguilles d'une horloge marquant six heures. Les observateurs dotés d'une vue normale voient une horloge indiquant exactement six heures.

FIGURE 3-19

Mire pour le test VT-12

BT.0500-03-19