RECOMMANDATION UIT-R BT.500-11

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

(Question UIT-R 211/11)

(1974-1978-1982-1986-1990-1992-1994-1995-1998-1998-2000-2002)

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 les différents laboratoires sur un certain nombre d'aspects de ces essais;

c) qu'il est important d'adopter des méthodes 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 méthodes recommandées pour les essais en laboratoire;

e) que l'introduction de nouvelles méthodes de traitement des signaux de télévision telles que le codage numérique et la réduction du débit binaire, les nouveaux types de signaux de télévision utilisant des composantes multiplexées dans le temps et, éventuellement, de nouveaux services tels que la télévision améliorée et la télévision à haute définition (TVHD) peuvent entraîner des modifications dans les méthodes d'évaluation subjective;

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 des opérations effectuées en amont,

recommande

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

2 que, à court terme, et malgré l'existence d'autres méthodes et la mise au point de nouvelles méthodes, les méthodes décrites aux § 4 et 5 de l'Annexe 1 de la présente Recommandation soient utilisées chaque fois que possible;

3 qu'étant donné qu'il est important de définir la base des évaluations 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;

4 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 présente Recommandation.

NOTE 1 – L'Annexe 1 fournit des informations sur les méthodes d'évaluation subjective servant à définir la qualité des systèmes de télévision.

NOTE 2 – L'Annexe 2 fournit une description des techniques statistiques de traitement des données recueillies au cours des essais subjectifs.

ANNEXE 1

Description des méthodes d'évaluation

1 Introduction

Les méthodes d'évaluation subjective 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. On les appelle 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. On les appelle couramment évaluations des dégradations.

Pour procéder à des évaluations subjectives appropriées, il faut d'abord choisir parmi les diverses options disponibles celles qui conviennent le mieux aux objectifs et aux circonstances du problème d'évaluation posé. A cet effet, à la suite des données générales du § 2, le § 3 donne quelques informations sur les problèmes d'évaluation que pose chaque méthode. Ensuite, les § 4 et 5 exposent en détail les deux principales méthodes recommandées. Enfin, le § 6 fournit des informations générales sur les autres méthodes étudiées.

La présente Annexe se limite à la description détaillée des méthodes d'évaluation. Le choix de la méthode la plus appropriée dépend cependant des objectifs du service étudié. Les procédures complètes d'évaluation d'applications spécifiques font donc l'objet d'autres Recommandations UIT-R.

2 Description générale

On trouvera dans cette section la description des conditions générales d'observation pour les évaluations subjectives. Les Recommandations connexes indiquent les conditions d'observation spécifiques pour des systèmes particuliers.

2.1 Conditions générales d'observation

On trouvera ci-après la description de différents environnements, pour lesquels les conditions d'observation diffèrent.

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.

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 un environnement proche de 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.

On trouvera plus loin une analyse de certains aspects du pouvoir de résolution et du contraste des écrans de contrôle (moniteurs).

2.1.1 Environnement laboratoire

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

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

a)	Rapport de la luminance de l'écran inactif à la luminance de crête:	≤ 0,02
b)	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:	≈ 0,01
c)	Brillance et contraste de la visualisation:	établi via PLUGE (voir les Recommandations UIT-R BT.814 et UIT-R BT.815)
d)	Angle maximal d'observation par rapport à la normale (cette valeur s'applique aux écrans cathodiques; les valeurs qui s'appliquent aux autres moyens de reproduction sont à l'étude):	30°
e)	Rapport de la luminance de l'arrière-plan, derrière le moniteur-image, à la luminance de crête de l'image:	≈ 0,15
f)	Chromaticité de l'arrière-plan:	D_{65}
g)	Eclairement de la salle dû à d'autres sources:	faible

2.1.2 Environnement domicile

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

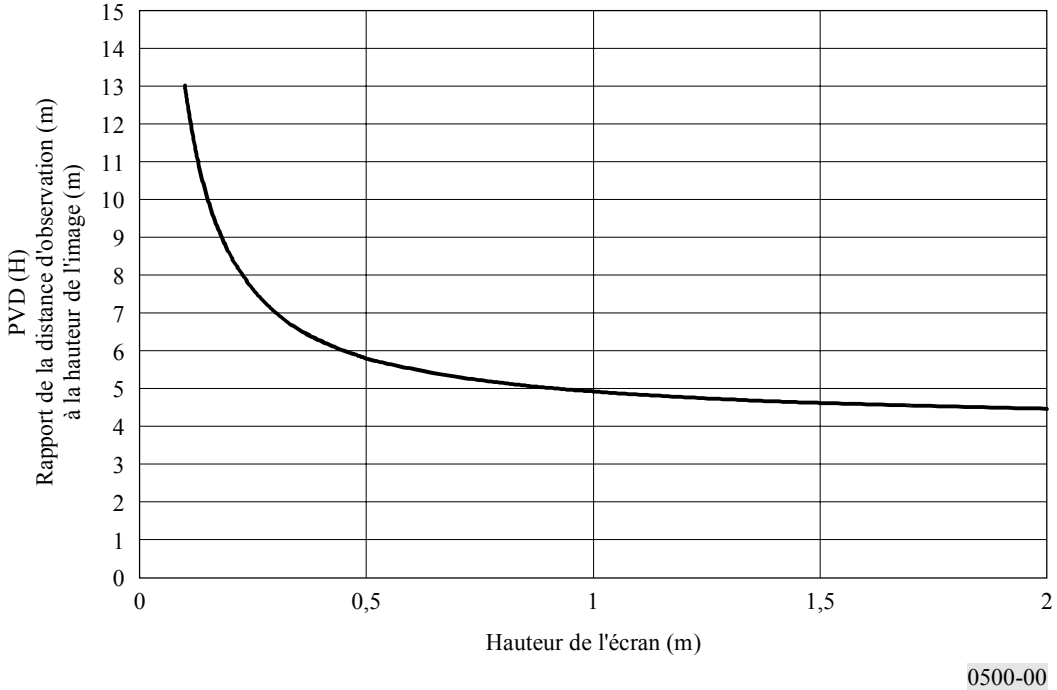
a)	Rapport de la luminance de l'écran inactif à la luminance de crête:	≤ 0,02 (voir le § 2.1.4)
b)	Brillance et contraste de la visualisation:	établi via PLUGE (voir les Recommandations UIT-R BT.818 et UIT-R BT.815)
c)	Angle maximal d'observation par rapport à la normale (cette valeur s'applique aux écrans cathodiques; les valeurs qui s'appliquent aux autres moyens de reproduction sont à l'étude):	30°
d)	Dimensions de l'écran pour un format 4/3:	ces dimensions d'écran doivent satisfaire aux règles relatives à la «distance d'observation préférée» (PVD, *preferred viewing distance*).
e)	Dimensions de l'écran pour un format 16/9:	ces dimensions d'écran doivent satisfaire aux règles relatives à la PVD

f)	Traitement du moniteur:	sans traitement numérique
g)	Pouvoir de résolution du moniteur:	voir le § 2.1.3
h)	Luminance de crête:	200 cd/m^2
i)	Luminance de l'environnement sur l'écran (La lumière incidente produite par l'environnement et tombant sur l'écran doit être mesurée perpendiculairement à l'écran):	200 lux

La distance d'observation et les dimensions de l'écran doivent être choisies de manière à être compatibles avec la PVD. Le tableau et le graphique ci-après donnent les valeurs de la PVD (en fonction des dimensions de l'écran). Ces chiffres pourraient être valables tant pour la télévision à définition normalisée (TVDN) que pour la TVDH, car on a relevé de très faibles différences.

Diagonale de l'écran (pouces)		**Hauteur de l'écran (*H*)**	**PVD**
Format 4/3	**Format 16/9**	**(m)**	**(*H*)**
12	15	0,18	9
15	18	0,23	8
20	24	0,30	7
29	36	0,45	6
60	73	0,91	5
> 100	> 120	> 1,53	3-4

PVD pour des images en mouvement



Ce tableau et ce graphique donnent des informations sur la PVD et les valeurs correspondantes des dimensions de l'écran, à adopter dans les Recommandations relatives à des applications spécifiques.

2.1.3 Pouvoir de résolution du moniteur

Le pouvoir de résolution des moniteurs professionnels, équipés de tubes cathodiques professionnels, est généralement conforme aux normes requises pour les évaluations subjectives, en ce qui concerne leur gamme de fonctionnement en luminance.

Tous les moniteurs ne sont pas capables d'atteindre une luminance de crête de 200 cd/m^2.

On pourrait envisager de vérifier et de rapporter les pouvoirs de résolution maximal et minimal (au centre et dans les angles de l'écran) pour la valeur de luminance utilisée.

Si des récepteurs de télévision grand public, avec tubes cathodiques grand public, sont utilisés pour les évaluations subjectives, on risque d'obtenir un pouvoir de résolution inadéquat, selon la valeur de la luminance.

Dans ce cas, il est fortement recommandé de vérifier et de rapporter les pouvoirs de résolution maximal et minimal (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 moniteurs ou des récepteurs de télévision grand public est un système à mire électronique avec balayage.

Le contrôle du pouvoir de résolution peut se faire par analyse visuelle. On estime que le seuil de vision se situe à –12/–20 dB. Le principal inconvénient de ce système est la présence d'un repliement du spectre, généré par le masque, qui rend l'évaluation difficile. Toutefois, la présence de ce repliement indique que le signal à vidéofréquence dépasse les limites fixées par le masque, ce qui a pour effet de sous-échantillonner le signal vidéo.

Il serait bon de poursuivre les études sur les essais relatifs à la définition des tubes cathodiques.

2.1.4 Contraste des moniteurs

Le contraste pourrait être fortement influencé par la luminance ambiante.

Il est rare que les tubes cathodiques des moniteurs professionnels mettent en œuvre des techniques capables d'améliorer leur contraste en présence d'une grande luminance ambiante. *Il est possible, par conséquent, que ces moniteurs ne satisfassent pas à la norme de contraste requise lorsqu'ils fonctionnent dans de telles conditions de luminance*.

Dans les tubes cathodiques grand public, on applique des techniques propres à améliorer le contraste en présence d'une grande luminance ambiante.

Pour calculer le contraste d'un tube cathodique donné, on a besoin de connaître le coefficient de réflexion, K, de son écran. Dans le cas le plus favorable, on a $K = 6\%$.

Dans un environnement du type diffus avec un éclairement I de 200 lux et $K = 6\%$, la formule suivante donne 3,82 cd/m^2 pour la luminance réfléchie des zones inactives de l'écran:

$$L_{réfléchie} = \frac{I}{\pi} K$$

Avec les valeurs indiquées plus haut, la luminance réfléchie (cd/m^2) est égale à près de 2% de l'éclairement incident (lux).

On considère que le tube cathodique ne donne pas de réflexions spéculaires sur la plaque de verre frontale, dont il est difficile de quantifier l'influence exacte sur le contraste, car celle-ci dépend grandement des conditions d'éclairage.

Aux § 2.1.1 et 2.1.2, le rapport de contraste *CR* s'exprime par:

$$CR = L_{min} / L_{max}$$

avec:

L_{min}: luminance des zones inactives dans l'éclairement ambiant (cd/m^2) (avec les valeurs indiquées: $L_{min} = L_{zones\ inactives} + L_{réfléchie}$ = 3,82 cd/m^2)

L_{max}: luminance des zones blanches dans l'éclairement ambiant (cd/m^2) (avec les valeurs indiquées: $L_{max} = L_{blanc} + L_{réfléchie}$ = 200 + 3,82 cd/m^2).

Avec ces valeurs, le calcul donne *CR* = 0,018, résultat très proche de la valeur 0,02 indiquée au a) des § 2.1.1.1 et 2.1.2.1.

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 et les séquences d'images mémorisées numériquement constituent les signaux source les plus facilement reproductibles et ce sont donc elles que l'on préfère. Elles peuvent être échangées entre différents laboratoires, cela afin d'obtenir de meilleures comparaisons entre systèmes. En outre, les formats de bande vidéo ou pour ordinateur peuvent convenir.

A court terme, les analyseurs de diapositives 35 mm sont préférables pour des 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.

Il faut 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.

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 indique les problèmes classiques d'évaluation et le matériel d'essai qui sert à les traiter.

TABLEAU 1

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 Appendices 1 et 2 à l'Annexe 1 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.

Un certain nombre d'organisations ont mis au point des images ou des séquences d'images d'essai fixes. A l'avenir, on espère traiter ce problème dans le cadre de l'UIT-R. Un matériel image spécifique est proposé dans les Recommandations relatives à l'évaluation des applications.

Les Appendices 1 et 2 à l'Annexe 1 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

Il faut 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 de télévision, ni des observateurs expérimentés (voir la Note 1). 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.

NOTE 1 – Selon des résultats préliminaires, des observateurs non spécialistes confrontés à une qualité de transmission et à des technologies de visualisation supérieures, donneraient des résultats plus critiques.

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 non spécialistes 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, mais, en attendant, les expérimentateurs devraient donner le plus de détails possible sur les caractéristiques des groupes d'évaluation qu'ils ont retenus afin d'étudier plus avant ce facteur. Ils pourraient par exemple donner 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.

2.6 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.

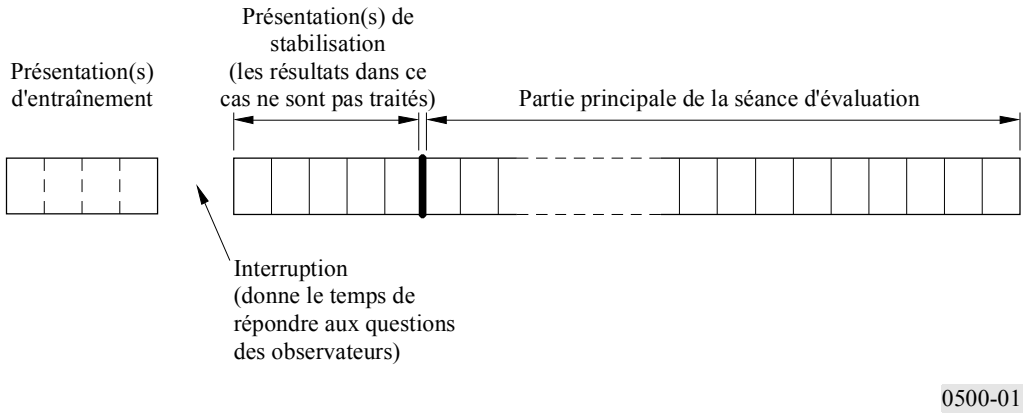
2.7 Séance d'évaluation

Une séance ne devra 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

Structure de la présentation d'une séance d'évaluation



2.8 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 2 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 de moniteur d'évaluation (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 – Etant 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. Le Tableau 2 indique les problèmes classiques d'évaluation et les méthodes qui servent à les traiter.

TABLEAU 2

Choix des méthodes d'essai

Problème d'évaluation	Méthode utilisée	Description
Mesurer la qualité des systèmes par rapport à une référence	Méthode à double stimulus utilisant une échelle de qualité continue (DSCQS)(1)	Rec. UIT-R BT.500, § 5
Mesurer la résistance des systèmes (c'est-à-dire vis-à-vis des défaillances)	Méthode à double stimulus utilisant une échelle de dégradation (DSIS)(1)	Rec. UIT-R BT.500, § 4
Mesure quantitative de la qualité des systèmes (quand on ne dispose d'aucune référence)	Méthode utilisant une échelle de rapports(2) ou une échelle catégorielle, à l'étude	Rapport UIT-R BT.1082
Comparer la qualité de plusieurs systèmes (quand on ne dispose d'aucune référence)	Méthode de comparaison directe ou méthode utilisant une échelle de rapports(2) ou une échelle catégorielle, à l'étude	Rapport UIT-R BT.1082
Identifier les paramètres qui permettent de déterminer les différences entre les systèmes et mesurer leur influence perceptuelle	Méthode à l'étude	Rapport UIT-R BT.1082
Définir le moment où une dégradation devient visible	Estimation du seuil par choix forcé ou méthode d'ajustement, à l'étude	Rapport UIT-R BT.1082
Déterminer si on remarque une différence entre systèmes	Méthode de choix forcé, à l'étude	Rapport UIT-R BT.1082
Mesurer la qualité du codage des images stéréoscopiques	Méthode à double stimulus utilisant une échelle de qualité continue (DSCQS)(3)	Rec. UIT-R BT.500, § 5
Mesurer la fidélité entre deux séquences vidéo dégradées	Méthode d'évaluation continue à double stimulus simultané (SDSCE)	Rec. UIT-R BT.500, § 6.4
Comparer différents mécanismes de protection contre les erreurs	Méthode d'évaluation continue à double stimulus simultané (SDSCE)	Rec. UIT-R BT.500, § 6.4

(1) La méthode DSCQS et la méthode DSIS ont donné lieu à un certain nombre d'études portant sur les effets contextuels. On a constaté que ces effets faussent quelque peu les résultats de la méthode avec échelle de dégradation. Des indications plus détaillées sont données dans l'Appendice 3 à l'Annexe 1.

(2) Certaines études donnent à penser que cette méthode est plus stable lorsqu'on dispose d'une large gamme de qualité.

(3) L'évaluation des images stéréoscopiques peut engendrer une grande fatigue. Il convient par conséquent de limiter la durée totale d'une séance d'essai à une valeur inférieure à 30 min.

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

4.1 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 au § 5.

La méthode à double stimulus (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. A 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.

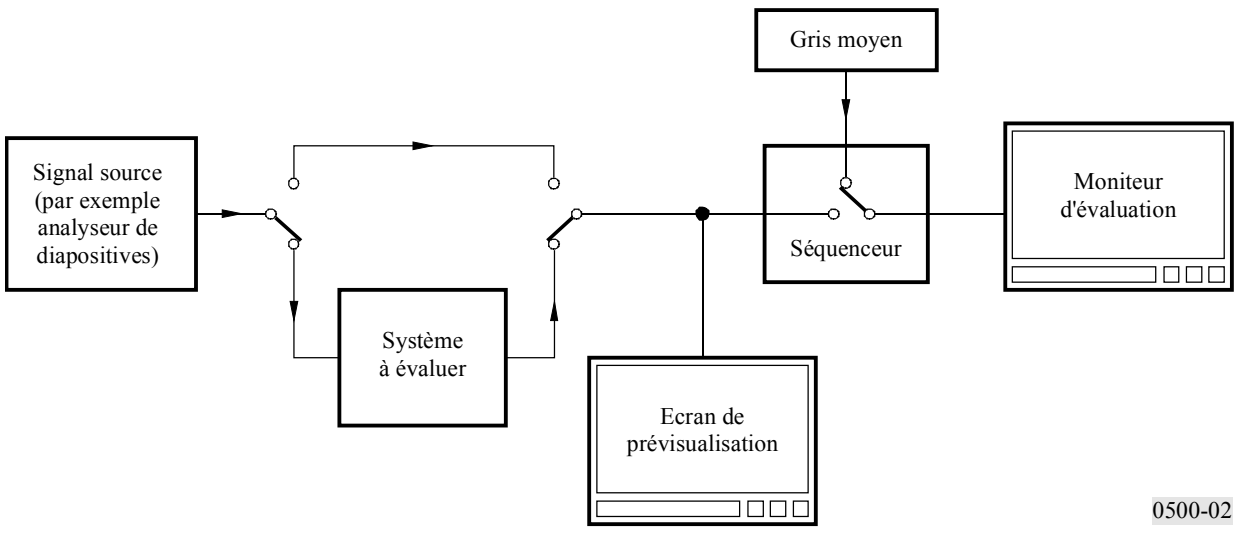
4.2 Mode opératoire général

Désigne la façon de définir ou de choisir, conformément au § 2, 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.

FIGURE 2

Configuration du système d'essai pour la méthode DSIS



On présente aux observateurs un moniteur 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 paires: 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.

4.3 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. 3a)).

Variante II: image ou la séquence de référence et l'image ou la séquence d'essai sont présentées deux fois (Fig. 3b)).

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.

4.4 Echelles 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.

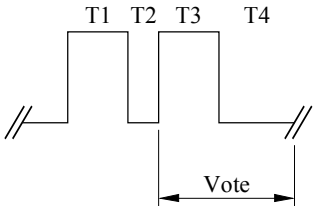
4.5 Introduction aux évaluations

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.

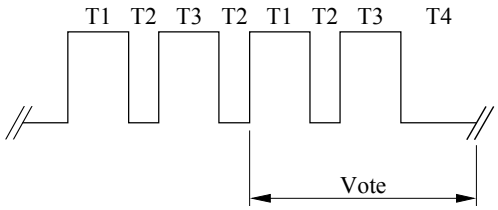
Les observateurs seront priés de regarder l'image pendant toute la durée de T1 et T3. Le vote sera autorisé uniquement pendant T4.

FIGURE 3

Structure de la présentation du matériel d'essai



a) Variante I



b) Variante II

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	Conditions 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.

0500-03

4.6 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'Appendice 2 à l'Annexe 1 donne des indications supplémentaires pour le choix des niveaux de dégradation.

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

5.1 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. A l'issue des séances, on calcule les moyennes pour chaque condition expérimentale et chaque image d'essai.

5.2 Mode opératoire général

Désigne la façon de définir ou de choisir, conformément au § 2, 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 § 4.6.

Le système d'essai aura la configuration indiquée à la Fig. 4.

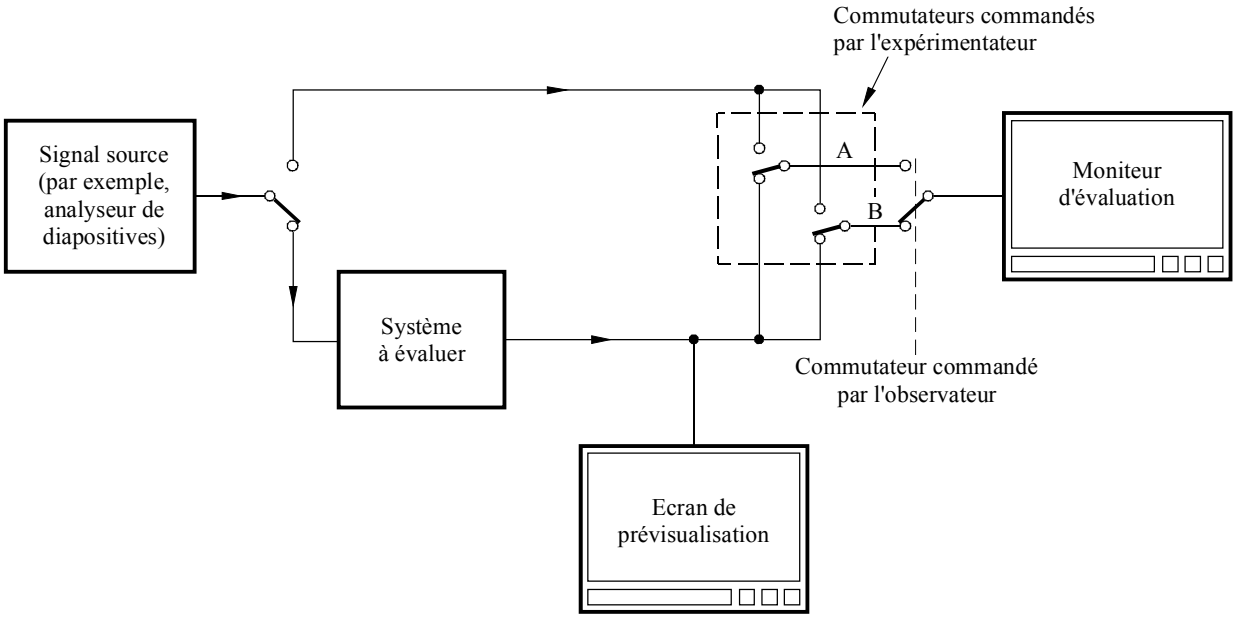
5.3 Présentation du matériel d'essai

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. 5.

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 4

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.

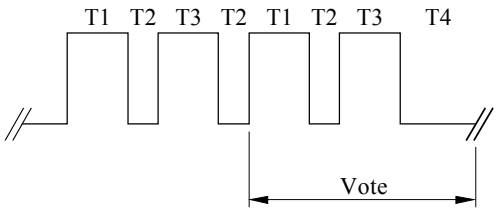
Variante I: l'observateur, habituellement seul, est autorisé à passer de la condition A à la B et inversement jusqu'à ce qu'il se soit fait une opinion sur chacune d'elles. Les trajets A et B reçoivent l'image de référence directe ou l'image qui est passée par le système à évaluer. L'image et le trajet 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 opinion sur chacune d'elles. Pour chaque présentation, les trajets A et B recoivent l'image comme dans la variante I ci-dessus. La stabilité des résultats obtenus avec cette variante, qui utilise une échelle de qualité limitée, est encore à l'étude.

0500-04

FIGURE 5

Structure de la présentation du matériel d'essai



Phases de la présentation:

T1 =	10 s	Séquence d'essai A
T2 =	3 s	Gris moyen produit par un niveau vidéo d'environ 200 mV
T3 =	10 s	Séquence d'essai B
T4 =	5-11 s	Gris moyen

0500-05

5.4 Echelle 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 Fig. 6 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.

5.5 Analyse des résultats

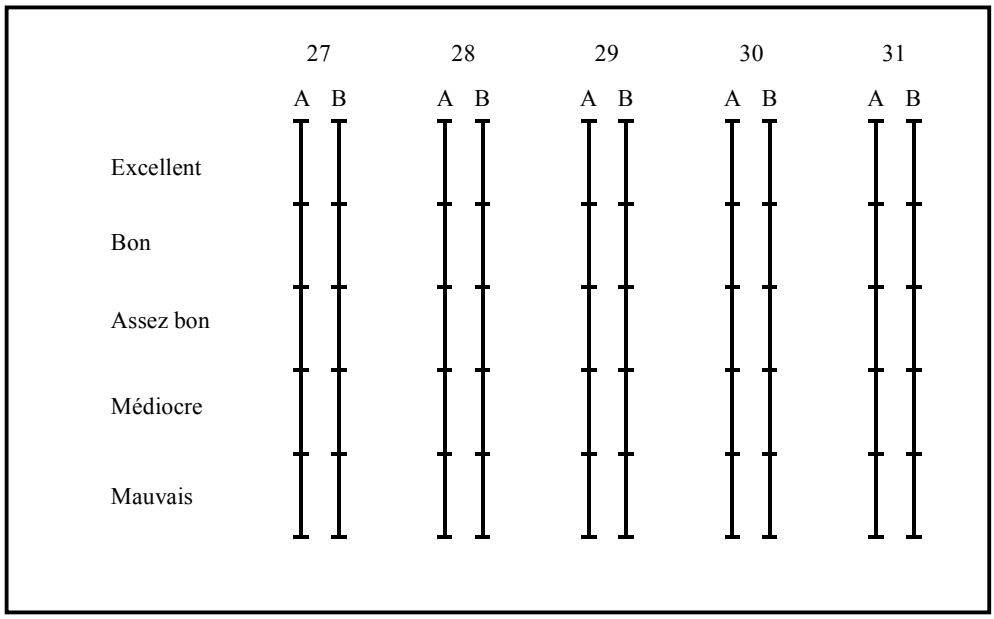
Pour les deux évaluations (image de référence et image d'essai) de chaque condition d'essai, les mesures de la longueur du trait indiqué sur la feuille de notation sont converties en notes normalisées comprises entre 0 et 100, puis, les différences de notation entre l'image de référence et l'image d'essai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

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'Appendice 1 à l'Annexe 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.

FIGURE 6

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



* Losqu'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 entâchée d'erreurs systématiques. Toutefois, les modalités de ces contrôles sont à l'étude.

0500-06

5.6 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.

6 Autres méthodes d'évaluation

Lorsque les conditions le permettent, on utilisera les méthodes à un seul stimulus et les méthodes de comparaison de stimulus.

6.1 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.

6.1.1 Mode opératoire général

Désigne la façon de définir ou de choisir, conformément au § 2, 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.

6.1.2 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.

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).

6.1.3 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); 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.

6.1.4 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.

6.1.4.1 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 3. Dans la surveillance de 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 3

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.

6.1.4.2 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.

6.1.4.3 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 3). 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).

6.1.4.4 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.

6.2 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.

6.2.1 Mode opératoire général

Désigne la façon de définir et de choisir, conformément au § 2, 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.

6.2.2 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.

6.2.3 Séance d'évaluation

L'évaluation fera intervenir soit un seul moniteur 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 moniteur 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 moniteurs 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.

6.2.4 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.

6.2.4.1 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 4.

TABLEAU 4

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.).

6.2.4.2 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 4). 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.

6.2.4.3 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.

6.3 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.

Une nouvelle technique, la technique SSCQE, décrite ci-après, a donc a été mise au point et testée.

6.3.1 Évaluation continue de la qualité globale

6.3.1.1 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.

6.3.1.2 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 min;

– *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.

6.3.1.3 Paramètres d'observation

Les conditions d'observation seront celles qui sont actuellement indiquées dans les Recommandations UIT-R BT.500, UIT-R BT.1128, UIT-R BT.1129 et UIT-R BT.710.

6.3.1.4 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 § 5.4.

6.3.1.5 Observateurs

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

6.3.1.6 Directives à l'intention des observateurs

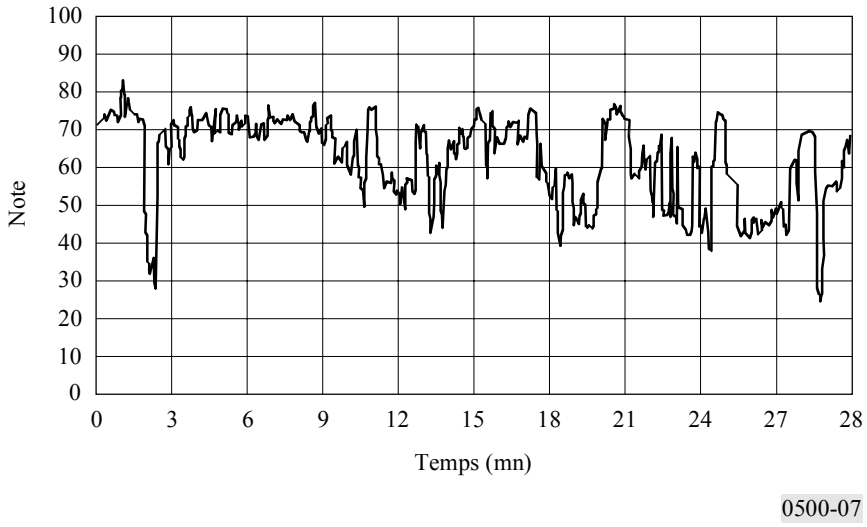
Dans le cas d'une évaluation de la qualité de service (avec séquence audio d'accompagnement), il sera demandé aux observateurs de juger la qualité d'ensemble et non la qualité vidéo seulement.

6.3.1.7 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. 7).

FIGURE 7

Condition de test: codex X/segment de programme: Z

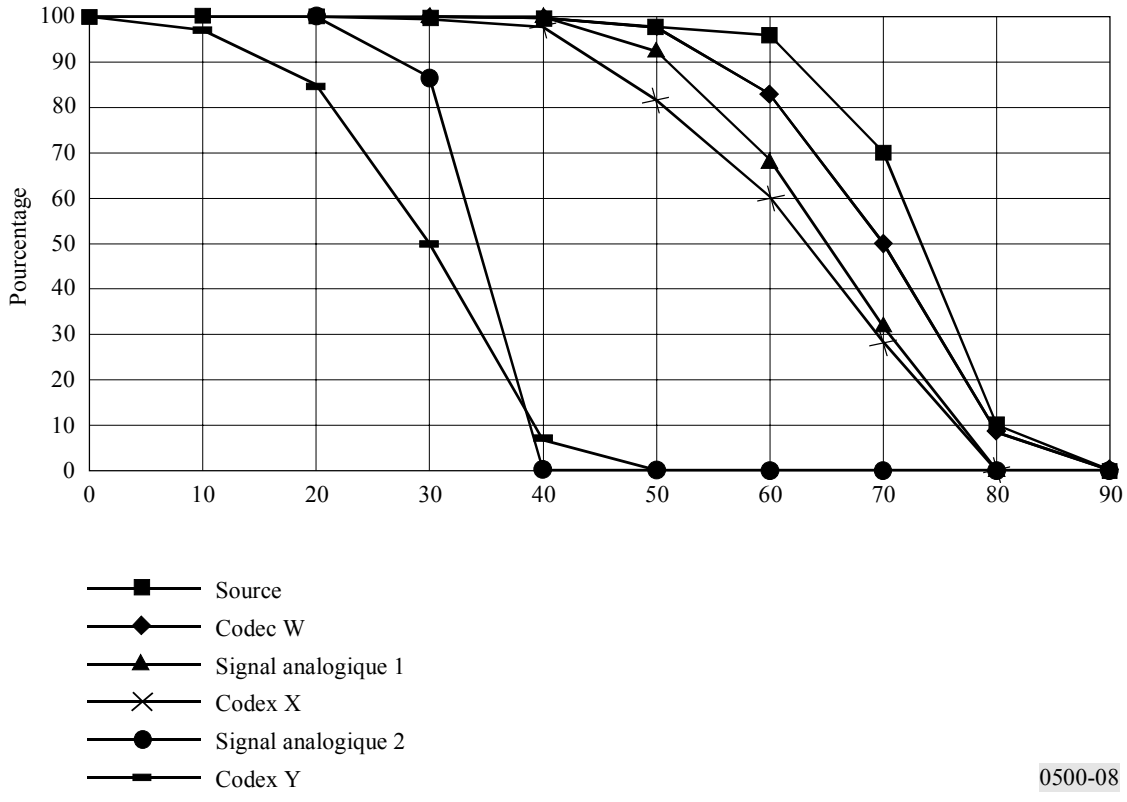


Toutefois, les différences de temps de réaction entre téléspectateurs peuvent influencer les résultats de l'évaluation si seule la moyenne sur un segment de programme est calculée. Des études destinées à évaluer l'incidence du temps de réaction sur la note de qualité obtenue sont actuellement en cours.

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. 8).

FIGURE 8

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



6.3.2 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.

6.4 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 méthodes 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.

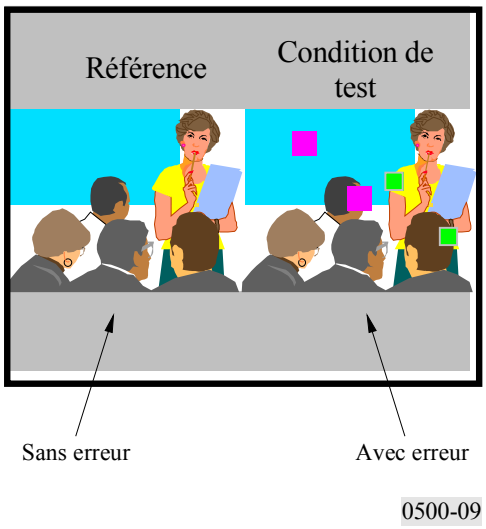
La nouvelle technique SDSCE décrite ci-dessous a donc été mise au point et testée.

6.4.1 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. 9).

FIGURE 9

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.

6.4.2 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.

6.4.3 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.

6.4.4 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. 10. Les résultats de cette étape peuvent être représentés par un diagramme temporel (Fig. 11).

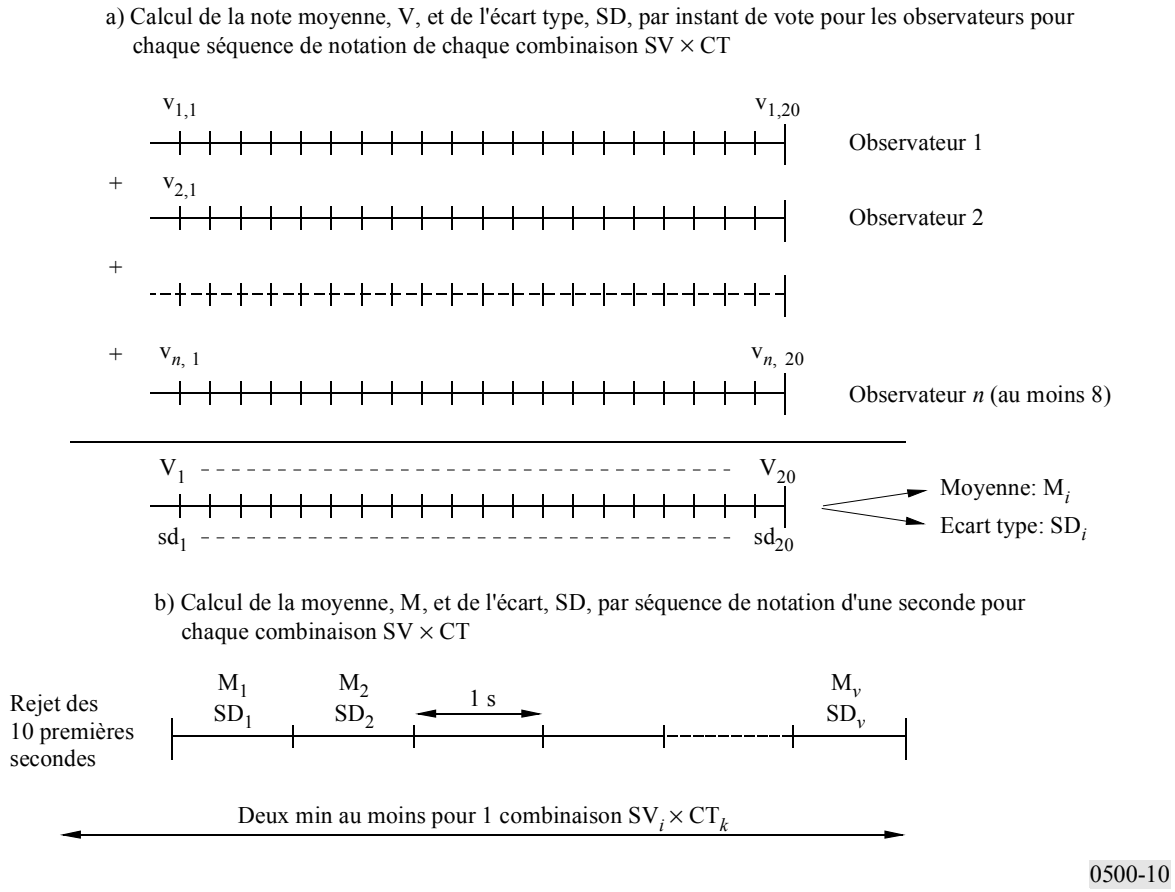
– On analyse la répartition statistique des moyennes calculées à l'étape précédente (c'est-à-dire celles correspondant à chaque segment de notation) et leur fréquence

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. 12. 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 10

Traitement des données



6.4.5 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 § 2.3.2 de l'Annexe 2 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 11

Diagramme temporel brut

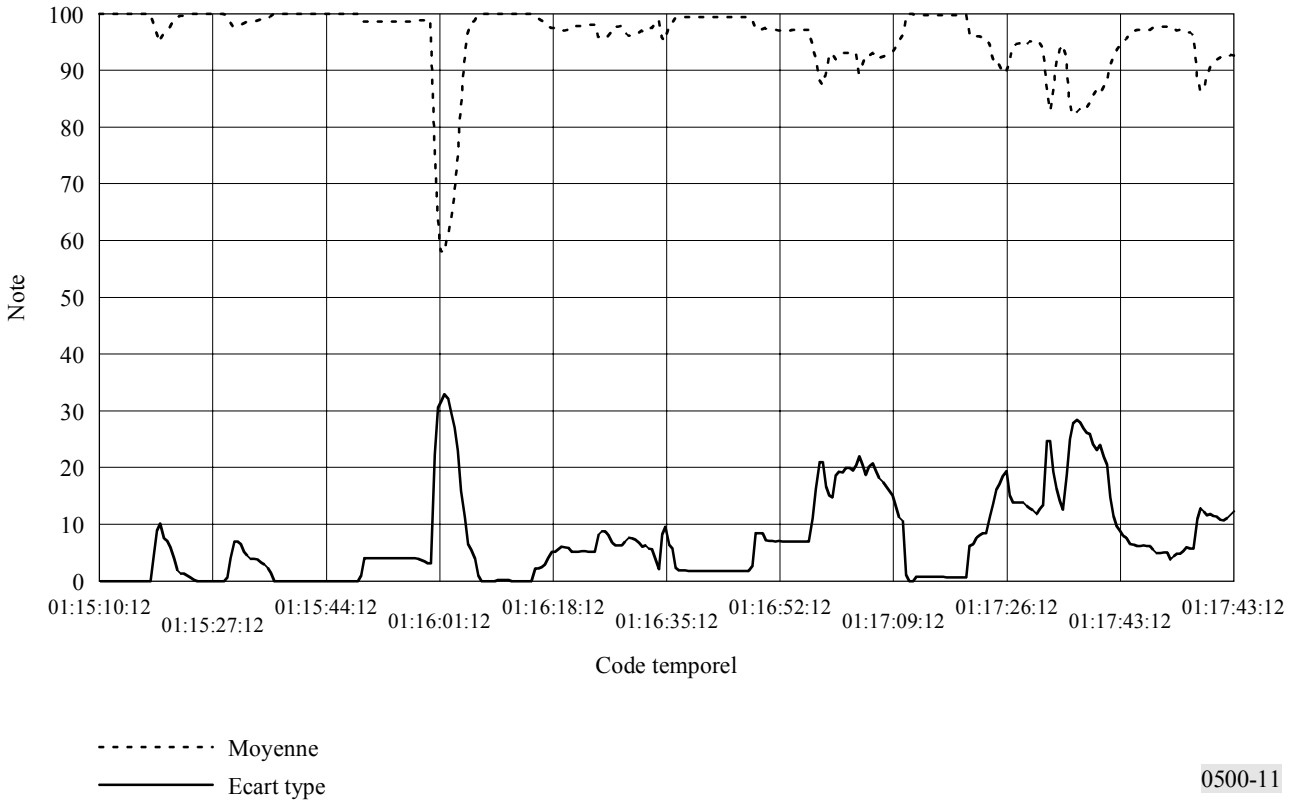
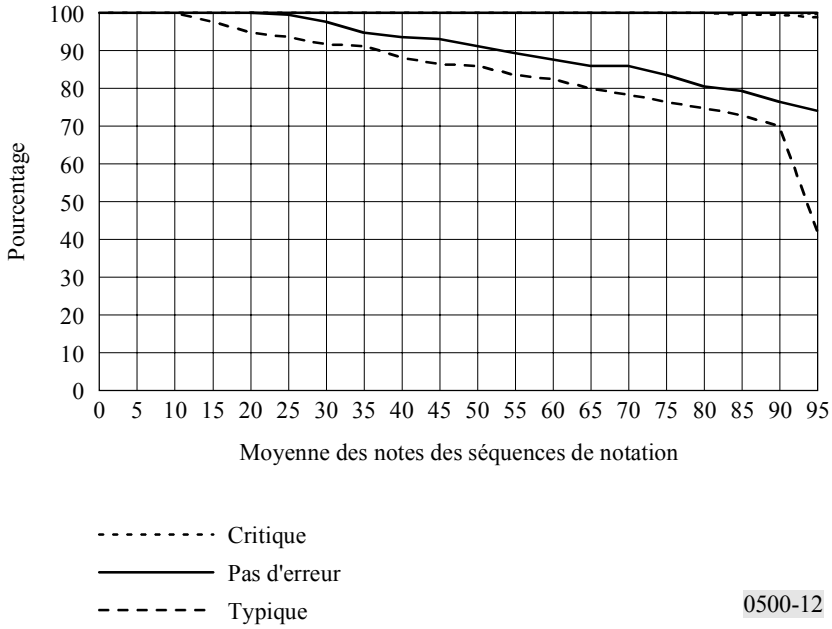


FIGURE 12

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



6.5 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 semble donc plus judicieux d'envisager des approches plus «complètes», c'est-à-dire d'utiliser plusieurs méthodes, ou une approche multidimensionnelle.

APPENDICE 1

À L'ANNEXE 1

Caractéristiques de dégradation du contenu de l'image

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.

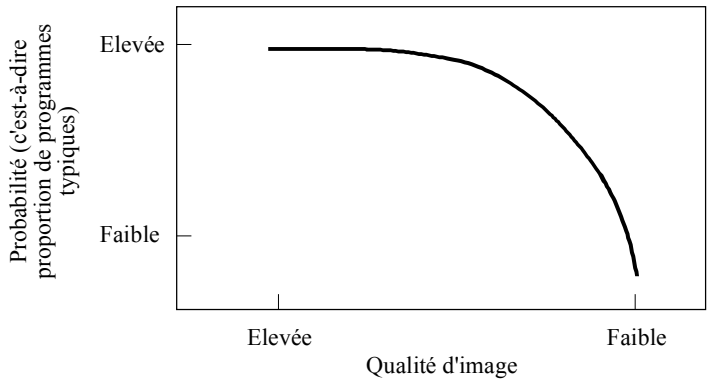
Cette caractéristique est particulièrement importante pour les systèmes dont la qualité ne se dégrade pas progressivement quand l'image devient de plus en plus critique. Par exemple, certains systèmes numériques et adaptatifs peuvent conserver une haute qualité sur toute une large gamme de types de programmes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

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. 13.

FIGURE 13

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



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

– *Etape 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.

– *Etape 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 Fig. 14 présente un exemple de ce genre de distribution.

– *Etape 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 Fig. 15 donne un exemple de cette fonction.

FIGURE 14

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

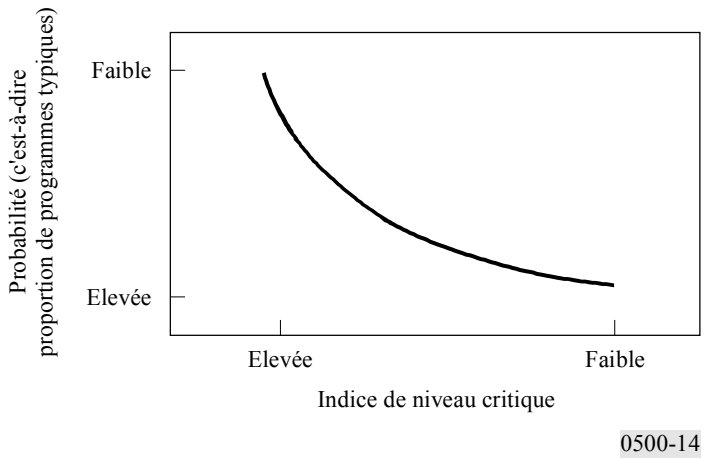
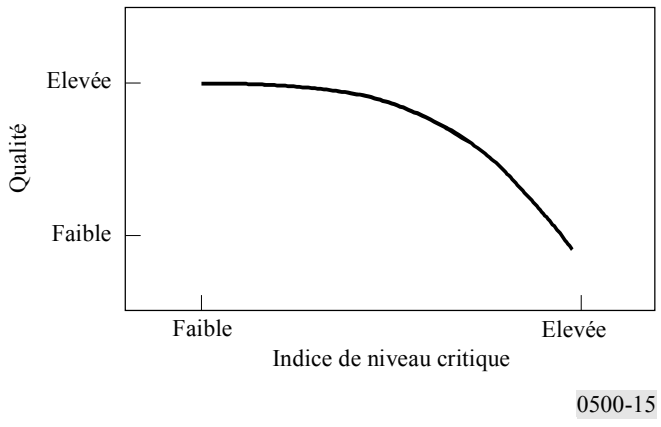


FIGURE 15

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



– *Etape 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. 13.

3 Utilisation de la caractéristique de dégradation

La caractéristique de défaillance, 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 résolution à la source, le débit binaire, la largeur de bande) au moment de sa conception pour l'adapter au mieux aux exigences d'un service;

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

– 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 Recommandation UIT-R BT.1129.

APPENDICE 2

À L'ANNEXE 1

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

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.

Le présent Appendice 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 caractéristiques de dégradation composite.

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'Appendice 1 à l'Annexe 1). On les combinera avec les résultats de l'analyse des conditions de transmission pour préparer l'étape finale de la méthode.

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.

4 Établissement des caractéristiques de dégradation composite

A 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 1, § 5). 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.

APPENDICE 3

À L'ANNEXE 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 2

Analyse et présentation des résultats

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 aux § 4, 5 et 6 de l'Annexe 1 à 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 multipoint. 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.

2 Méthodes communes d'analyse

Les tests effectués conformément aux principes des méthodes décrites dans l'Annexe 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.

2.1 Calcul des notes moyennes

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

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

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, $\overline{u}_j$ et $\overline{u}_k$, pour chaque condition et séquence/image de test.

2.2 Calcul de l'intervalle de confiance

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:

$$\left[\overline{u}_{jkr} - \delta_{jkr},\ \overline{u}_{jkr} + \delta_{jkr}\right]$$

où:

$$\delta_{jkr} = 1{,}96 \frac{S_{jkr}}{\sqrt{N}} \tag{2}$$

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

$$S_{jkr} = \sqrt{\sum_{i=1}^{N} \frac{(\overline{u}_{jkr} - u_{ijkr})^2}{(N-1)}} \tag{3}$$

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.

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.

2.3 Sélection des observateurs

2.3.1 Sélection 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 $\overline{u}_{jkr}$ 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 1).

NOTE 1 – 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 $\overline{u}_{jkr}$, l'écart type S_{jkr}, et le coefficient de kurtosis β_{2jkr}, où β_{2jkr} est donné par:

$$\beta_{2jkr} = \frac{m_4}{(m_2)^2} \quad \text{avec} \quad m_x = \frac{\sum_{i=1}^{N} (u_{ijkr} - \overline{u}_{ijkr})^x}{N} \tag{4}$$

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

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

sinon:

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

si $u_{ijkr} \le \overline{u}_{jkr} - \sqrt{20}\,S_{jkr}$ 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).

2.3.2 Sélection 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.

Etape 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 $\overline{u}_{jklr}$, l'écart type S_{jklr}, et le coefficient de β_{2jklr}, où β_{2jklr} est donné par:

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

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

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

sinon:

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

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

Si $\dfrac{P_i}{J \cdot K \cdot L \cdot R} > X\%$ ou $\dfrac{Q_i}{J \cdot K \cdot L \cdot R} > X\%$ 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 Fig. 17 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.

Etape 2: Détection des inversions de note locales

Dans l'Etape 2, la détection se fonde également sur les formules de sélection données dans l'Annexe 2 de la présente Recommandation 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, $\overline{u}_{klr}$ est définie par:

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

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

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

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:

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

Pour chaque fenêtre temporelle de chaque configuration de test, on calcule la moyenne $\bar{u}^*_{jklr}$, l'écart type S^*_{jklr}, et le coefficient $\beta_2{}^*_{jklr}$, ce dernier coefficient étant donné par:

$$\beta_2{}^*_{jklr} = \frac{m_4}{(m_2)^2} \quad \text{avec} \quad m_x = \frac{\sum_{n=1}^{N} (u^*_{njklr})^x}{N}$$

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

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

Si $\dfrac{P^*_i + Q^*_i}{J \cdot K \cdot L \cdot R} > Y$ 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.

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 $\overline{u}$ ($\overline{u}$ calculé comme indiqué au § 2.1), la méthode suivante peut être utile; elle consiste à rechercher une correspondance simple entre $\overline{u}$ et le paramètre dégradation.

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 $\overline{u}$ peut se faire comme suit:

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

$$p = (\overline{u} - u_{min}) / (u_{max} - u_{min}) \tag{5}$$

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:

$$p = 1 / [1 + \exp (D - D_M)\, G] \tag{6}$$

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:

$$I = (1/p - 1) \tag{7}$$

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

$$I = \exp (D - D_M)\, G \tag{8}$$

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

$$\log_e I = (D - D_M)\, G \tag{9}$$

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:

$$S = \frac{D_M - D}{\log_e I} = \frac{1}{G} \tag{10}$$

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.

3.2 Approximation par une fonction non symétrique

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 (8) doit être remplacée par:

$$I = (d / d_M)^{1/G} \tag{11}$$

ce qui donne pour la relation (6):

$$p = 1 \Big/ \left[1 + (d / d_M)^{1/G} \right] \tag{12}$$

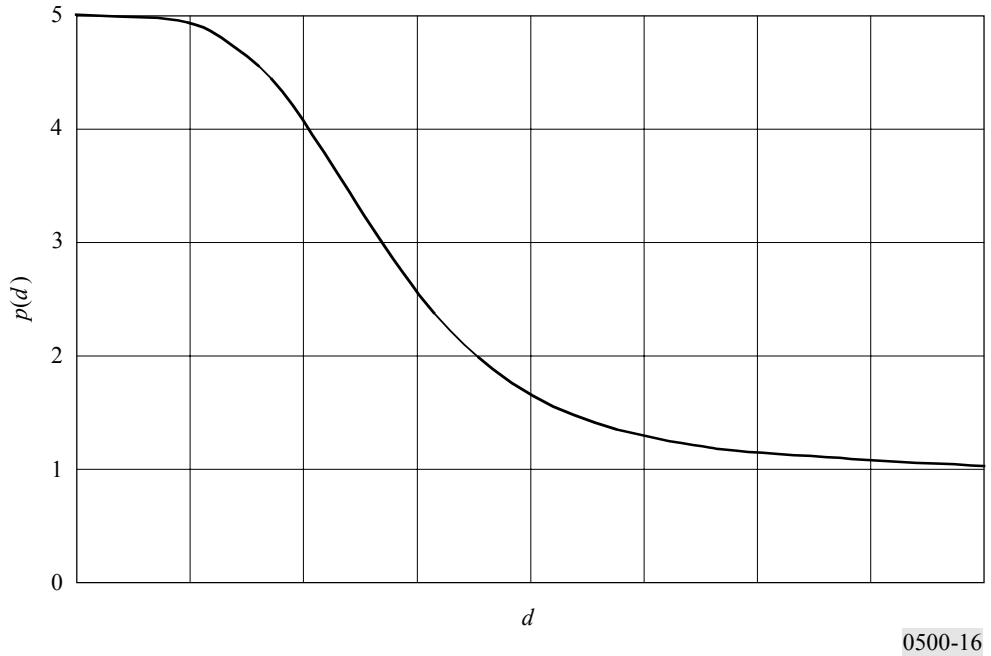
Cette fonction constitue une approximation non symétrique de la fonction logistique.

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 Fig. 16 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 16

Approximation non symétrique



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 $\overline{u}_0 < u_{max}$).

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 5).

TABLEAU 5

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:

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

$$C = \frac{\overline{u} - u_{0\,min}}{u_{0\,max} - u_{0\,min}} \frac{u_{max} - u_{mid}}{u_{0\,max} - u_{mid}} + \frac{u_{0\,max} - \overline{u}}{u_{0\,max} - u_{0\,min}} \frac{u_{min} - u_{mid}}{u_{0\,min} - u_{mid}}$$

avec:

u_{corr} : note corrigée

$\overline{u}$: 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.

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.

3.4 Incorporation de l'aspect fiabilité dans les graphiques

A 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);

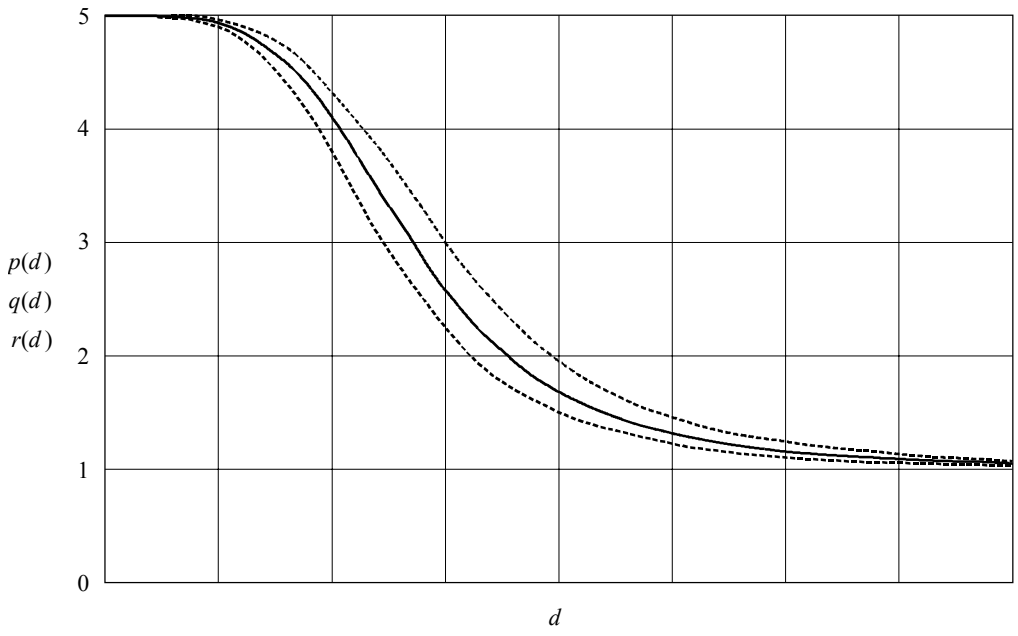
– série de notes moyennes;

– 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. 17). On obtient ainsi une estimation de la zone de confiance continue à 95%.

FIGURE 17

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

0500-17

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.

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.

ANNEXE 3

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 Recommandation UIT-R BT.500.

Il se compose de fichiers de texte, dont la structure est illustrée dans les Tableaux 6 et 7. 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 6

Format de fichier de données pour l'identification des résultats

Format et syntaxe du fichier d'identification	Commentaires
[Structure du test]↵	[Identificateur de section]
Type = *«DSCQS» ou «DSIS I», «DSIS II», etc.*	Identification de la méthodologie Rec. UIT-R BT.500 utilisée
Nombre de séances = *1 ≤entier≤ x*↵	Nombre de séances(1) par test
Minimum de l'échelle = *entier*↵	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 de contrôle = *entier*↵	Longueur de la diagonale(pouces)
fabricant du moniteur et modèle = *chaîne de caractères*↵	
[RÉSULTATS] ↵	[Identificateur de section]
Nombre de résultats = *1 ≤ entier ≤ y*↵	Nombre de fichiers résultats(1) pris en considération
Résultat(*j*).Nom de fichier(s) = *chaîne de caractères*.DAT↵	Nom de fichier complet.DAT (voir le Tableau 7), y compris le chemin
....	
Résultat(*j*).Nom = *chaîne de caractères*↵	Nom du fichier résultats habituel
Résultat(*j*).Laboratoire = *chaîne de caractères*↵	Identification du laboratoire effectuant le test
Résultat(*j*).Nombre d'observateurs = *1 ≤ entier ≤ N*↵	Nombre total d'observateurs
Résultat(*j*).Initialisation = *«Oui» ou «Non»* ↵	Indique si les notes recueillies pendant l'initialisation sont incluses dans le fichier DAT joint
[Résultat(*j*).Séance(*i*).Observateurs] ↵	[Identificateur de section]
O(*k*).Premier Nom = *chaîne de caractères*↵	Identification de l'observateur
O(*k*).Dernier Nom = *chaîne de caractères*↵	
O(*k*).Sexe = *«F» ou «M»* ↵	Facultatif
O(*k*).Âge = *entier*↵	Facultatif
O(*k*). Activité professionnelle = *chaîne de caractères*↵	Principaux groupes socio-économiques (ouvriers, étudiants, par exemple)
O(*k*).Distance = *entier*↵	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 fichiers.DAT qui seront remis pour chaque exécution.

TABLEAU 7

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

Format et syntaxe du fichier nom de fichier.DAT	Commentaires
entier entier entier.....↵ entier entier entier.....↵	Un fichier de données brutes DAT se compose de valeurs de notation séparées par un espace. 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 6 par fichier(s) résultat(s) (*j*)(1)

(1) Voir le renvoi (1) du Tableau 6.