

MESURES OBJECTIVES ET SUBJECTIVES DE LA QUALITE DES
IMAGES DE TELEVISION EN FONCTION DES DEGRADATIONS
IMPORTANTES QUI RESULTENT DU CODAGE NUMERIQUE
DES SIGNAUX DE TELEVISION

(Question 3/11 et Programme d'études 3B/11)

(1990)

1. Introduction

L'évaluation des dégradations dues au codage est devenue d'une importance cruciale avec l'utilisation accrue du codage numérique et de la transmission à débit binaire réduit. La compréhension des méthodes d'évaluation est importante non seulement pour étudier la qualité de codage des nouveaux équipements, mais aussi pour interpréter les résultats des mesures faites sur des appareils de ce genre, ainsi que les définitions des objectifs de qualité. De plus, les codecs numériques et toutes les opérations numériques adaptatives ou non linéaires ne peuvent pas être pleinement caractérisés au moyen des signaux d'essai et des mires traditionnels de la télévision.

Les études dans le cadre de la Question 3/11 et du Programme d'études 3B/11 indiquent qu'il est souhaitable d'établir des relations entre les mesures objectives sur les signaux dégradés par le codage numérique et les évaluations subjectives de la qualité de l'image ainsi obtenue. Le présent rapport rend compte des progrès accomplis dans ce sens, mais le but est de plus en plus difficile à atteindre avec l'accroissement de complexité des codecs.

La Recommandation 500 décrit les méthodes d'évaluation subjective de la qualité et des dégradations d'une image de télévision de définition normale et la Recommandation 710 celles de la TVHD. Le présent Rapport est consacré à l'application de ces méthodes aux codecs de télévision.

On a acquis récemment une grande expérience dans l'évaluation subjective des codecs de haute qualité pour la télévision en composantes 4:2:2 à 34, 45 et 140 Mbit/s [CCIR, 1986-90a]. On a étudié dans ces expériences les performances des codecs, des points de vue de la qualité intrinsèque de l'image décodée, de la qualité après des post-traitements de studio (incrustations et ralenti) de l'image décodée et des dégradations dues à différents taux d'erreur dans la voie. Le présent Rapport est en partie basé sur ces expériences.

Pour les applications de distribution la qualité peut être définie en termes de jugement subjectif de la part d'observateurs. En théorie, la qualité d'un codec prévue pour des applications de contribution pourrait donc être définie en termes de caractéristiques objectives ou subjectives car le signal qu'il produit n'est pas destiné au visionnage immédiat, mais à la post-production, à l'enregistrement et/ou au codage pour transmission ultérieure. Du fait de la difficulté de définir la qualité pour une série d'opérations de post-production différentes, le principe choisi a consisté à définir les performances d'une chaîne d'appareils, y compris pour la fonction de post-production, considérée comme représentative d'une application de contribution réelle. Cette chaîne pourrait être constituée d'un codec, suivi par un dispositif de post-production (ou par un autre codec dans le cas de l'évaluation de la qualité intrinsèque de la contribution), puis par un autre codec avant présentation du signal à l'observateur. L'adoption de cette stratégie pour les codecs destinés aux applications de contribution a pour conséquence que les procédures de mesure décrites dans le présent rapport peuvent aussi être utilisées pour les évaluer.

On insiste dans tout le rapport sur l'importance du choix de séquences d'essai critiques, surtout des scènes naturelles, et l'on donne quelques indications sur la manière dont ces séquences peuvent être produites ou choisies.

2. Classification des codes numériques

La fonction du codage numérique est de réduire le débit binaire nécessaire pour représenter une séquence d'images tout en garantissant une perte de qualité d'image minimale. Le matériel numérique y parvient en éliminant de l'image le plus possible d'éléments statistiquement redondants (cette première étape théorique n'introduit donc aucune perte de qualité). Si l'on doit réduire davantage le débit binaire, il faut introduire une certaine distorsion dans l'image, mais l'un des objectifs dans la conception d'un codec est de masquer cette distorsion en exploitant certaines insensibilités de perception du système visuel humain.

Il est commode de répartir les codecs entre deux classes regroupant ceux utilisant un codage à longueur de mot fixe et ceux où cette longueur est variable (on trouvera respectivement les définitions de ces termes aux § 3.1 et 3.2). La seconde classe a un meilleur rendement et est plus complexe; elle comprend tous les systèmes récemment proposés pour le codage de la vidéo 4:2:2 entre 30 et 45 Mbit/s. La première classe est cependant suffisante pour réduire la vidéo 4:2:2 à 140 Mbit/s, tout en conservant la qualité exigée pour les applications de contribution. Il est aussi utile de subdiviser encore ces classes en considérant les codecs intratrames (ou spatiaux) et ceux interimages (ou intertrames) comportant des mémoires d'image (ou de trame) leur permettant d'exploiter la redondance qui existe entre images (ou trames) consécutives.

Il commence à apparaître une troisième classe de codecs utilisant un codage à mots de longueur variable, mais conçus pour des réseaux à débit binaire variable. Ces codecs peuvent en principe conserver à l'image décodée une qualité constante jusqu'au maximum de charge du réseau. Les essais de qualité de tels codecs devraient tenir compte de la nature du réseau utilisé, ainsi que des statistiques des données injectées par tous ses usagers et ils restent à étudier.

3. Evaluation objective des codecs en termes de dégradation perçue de la qualité d'image

3.1 Codecs à mots de longueur fixe

Dans les codecs à mot de longueur fixe, on utilise un nombre invariable de bits pour représenter un nombre donné d'échantillons de l'image initiale. Par exemple, dans les codecs MIC ou MICD, à mots de longueur fixe, un nombre invariable de bits est affecté à la représentation de chaque échantillon de l'image et dans ceux du type à transformation ou à quantification vectorielle à mots de longueur fixe, un nombre invariable de bits est attribué à chaque bloc d'échantillons de l'image.

3.1.1 Méthodes basées sur l'utilisation de signaux d'essai artificiels

Dans ces codecs, la dégradation introduite dans chacun des échantillons d'image reçus dépend de la valeur de ses voisins, soit dans la même trame (codecs intratrames), soit dans la même trame et dans la précédente (codecs intertrames). Il est donc possible de provoquer artificiellement au moyen de signaux d'essai numérique bi ou tridimensionnels les dégradations caractéristiques du codage numérique appliqué à l'image.

Certains des facteurs de dégradation ont reçu des noms, comme les faux contours, le bruit granulaire, le flou, les effets de blocs, etc., qui correspondent à la manière dont ils sont perçus par les observateurs. Après avoir provoqué ces distorsions on peut mesurer objectivement leur amplitude et en mettant à profit l'expérience acquise au cours des évaluations subjectives, les résultats peuvent être mis en relation assez quantitative avec la qualité du codec. On donne des exemples de telles mesures dans [Kobayashi, 1977] pour les codecs intratrames et dans [Hishiyama & Inoue, 1984] pour ceux du type intertrames. Il peut être difficile d'établir une relation entre les facteurs de dégradation et leur interprétation par les observateurs dans les systèmes de codage intertrames ou dans ceux utilisant un traitement adaptatif car cette relation peut varier à tout moment en fonction des mouvements ou de l'adaptation de l'algorithme de codage. On présente dans [CCIR, 1982-86] une méthode de classement adaptée aux cas de ce genre. Dans cette méthode, on utilise d'abord pour les essais subjectifs des échelles établies à partir d'adjectifs de sens opposé (technique de la différence sémantique), puis les résultats sont analysés par analyse de la composante principale pour en extraire les facteurs de dégradation de la qualité d'image. Les résultats du classement peuvent être testés en appliquant une analyse de régression multiple qui met en relation les facteurs de dégradation et les jugements subjectifs. Le Tableau I présente une liste de facteurs de dégradation de la qualité d'image.

Ces méthodes semblent convenir pour l'évaluation des codecs et constituer aussi des outils de conception d'un codec, mais elles sont difficiles à mettre en relation avec les performances de l'appareil en face d'images réelles, et ceci pour les raisons suivantes:

- la composition complexe d'une séquence d'images réelles ne peut pas être représentée de manière satisfaisante par un nombre raisonnable de signaux d'essai artificiels;
- les dégradations peuvent être de nombreux types et être difficiles à classer du fait de leur nature subtile (par exemple, une distorsion donnée peut n'être visible que dans les parties texturées d'une image en mouvement d'une façon particulière);



- il risque d'être difficile de définir des mesures objectives significatives des dégradations (par exemple, pour le rendu du mouvement). Signalons que la durée de ces mesures devrait correspondre à la fenêtre d'observation dont on dispose pendant la présentation des essais subjectifs.

TABLEAU I

Exemples de facteurs de dégradation de la qualité dans les systèmes numériques et mesures physiques correspondantes (unités)

Facteur de dégradation de la qualité d'image	Mesure physique
Flou	Temps de montée de la réponse impulsionnelle
Flottement de contours	Gigue dans la réponse impulsionnelle
Faux contours	S_{p-p} au minimum c-c d'erreur de quantification
Bruit granulaire	Rapport signal/bruit analogique équivalent exprimé en termes de $S_{p-p} N_{eff}$.
Effet de vitre sale	Amplitude maximale du bruit
Flou temporaire	Temps de montée d'une transition mobile
Instabilité	Différence de position des transitions mobiles entre images ou trames
Dégradation en mosaïque	A étudier
Erreurs binaires	A étudier

3.1.2 Méthodes basées sur des images naturelles et des erreurs de codage

On peut considérer les séquences d'images naturelles comme composées d'un certain nombre de régions différentes du point de vue de leur nature et de l'influence qu'elles exercent sur les codecs à longueur de mot fixe. La nature de la séquence aura donc un effet significatif sur la qualité perçue par l'observateur [Roufs et autres, 1989]. Il est aussi possible, en cas de comparaison entre deux codecs, que ce soit la séquence en question qui détermine quel est celui des codecs paraissant être le meilleur. Ces faits mettent en lumière l'importance du choix des images employées dans les évaluations subjectives (voir le § 9), mais ils démontrent aussi que la mesure objective d'un codec donné doit prendre en compte la nature de l'image si l'on veut établir une relation entre les résultats de cette mesure et ceux de l'évaluation subjective.

Les formes les plus courantes de mesure objective de la qualité sont basées sur les erreurs de codage d'un codec, c'est-à-dire sur la différence entre la séquence d'images appliquée à son entrée et celle obtenue en sortie après décodage. Ce signal de différence (souvent d'amplitude) peut lui-même être affiché et il constitue une aide utile pour les spécialistes de la conception des codecs. Il ne doit cependant pas être utilisé dans les essais subjectifs.

3.1.3 Méthodes basées sur l'erreur quadratique moyenne normalisée

L'erreur quadratique moyenne de codage est fréquemment utilisée dans la mesure objective de la qualité des images décodées. Il s'agit de la moyenne sur l'ensemble des échantillons d'une séquence d'images des carrés de l'erreur de codage sur les échantillons; sa valeur est généralement normalisée par rapport au carré de la largeur de la dynamique des échantillons d'image. On donne quelquefois l'erreur quadratique moyenne normalisée (EQMN) comme facteur de bruit de codage, sous la forme $-10 \log (EQMN)$. La popularité de la méthode EQMN tient à sa simplicité mathématique, mais le résultat ne doit être considéré qu'avec prudence comme une mesure de la qualité après décodage. Par exemple, elle est incapable de faire la distinction entre un petit nombre d'importantes erreurs de codage (gênantes pour l'observateur) et un grand nombre d'erreurs de faible amplitude susceptibles de ne pas être perçues. On a tenté de pondérer le signal d'erreur de codage par une opération logarithmique avant l'évaluation de EQMN, grâce à un filtre représentant un modèle de la vue et on a ainsi obtenu une meilleure corrélation avec le résultat des évaluations subjectives. La EQMN est un bon outil pour la mise au point des codecs où l'on doit souvent comparer des méthodes de codage très semblables, c'est-à-dire utilisant des variantes mineures d'un même algorithme et où l'on peut admettre que le traitement des erreurs est identique.

3.1.4 Méthodes basées sur des modèles de la vue

La sensibilité du système visuel humain aux erreurs de codage dans une région donnée de l'image est fortement influencée par les caractéristiques de l'image elle-même dans cette région. Son incapacité à tenir compte de ce fait est un des principaux points faibles de la méthode EQMN. On peut donner l'exemple suivant de cette influence: on sait que la sensibilité d'un observateur au bruit dû à une erreur de codage est réduite quand le spectre de ce bruit coïncide approximativement avec celui de l'image "support". Ces propriétés du système visuel sont celles qui sont exploitées dans les codecs où l'on utilise des expériences subjectives ou des données psycho-visuelles pour optimiser les paramètres du système.

Pour améliorer la correspondance entre les mesures objectives de la qualité d'image et les jugements portés par les observateurs humains, il faut créer un modèle de la vue capable d'interpréter les erreurs de codage locales en tenant compte de l'image support et de combiner toutes ces évaluations locales pour parvenir à une note de qualité globale. Ce principe est applicable aux codecs à longueur de mot fixe ou variable et il est examiné au § 3.2.3.

3.2 Codecs à longueur de mot variable

Dans les codecs de télévision où le débit binaire de l'image de source doit être divisé par un facteur dépassant deux, environ, on emploie des méthodes basées sur le codage à longueur de mot variable. Ces codecs ont un meilleur rendement car ils possèdent une souplesse leur permettant d'affecter dynamiquement les bits de codage aux parties de l'image où ils permettent le mieux de maintenir la qualité de l'image décodée. Il existe plusieurs manières d'y parvenir, sans nécessairement employer un codage à entropie.

3.2.1 Méthodes basées sur l'utilisation de signaux artificiels

Du fait de la souplesse de ces codecs, les dégradations qu'ils peuvent causer à un échantillon codé dépend non seulement de la valeur des échantillons voisins, mais aussi de l'évolution dans le temps de ces échantillons, jusqu'à une image antérieure ou plus. Il en résulte que pour les codecs intra ou intertrames à mots de longueur variable, si l'on essayait de définir leurs caractéristiques en tentant de provoquer des distorsions locales au moyen de signaux d'essai également locaux et en mesurant ceux-ci le résultat n'aurait pas de sens. Toutefois, si l'on peut maintenir artificiellement les modes d'adaptation d'un codec à longueur de mot variable (ce qui signifie qu'on doit pouvoir accéder à ses mécanismes internes), chaque mode peut être caractérisé séparément. On peut alors se servir des informations sur la commutation des modes d'adaptation du codec, lorsque celui-ci est utilisé avec des scènes naturelles, afin de déterminer objectivement sa qualité de fonctionnement.

Il est possible d'imaginer des séquences d'essai artificielles en mouvement capables d'amener le codec au point où il provoque une distorsion visible, mais même si l'on pouvait définir des mesures objectives caractérisant ces distorsions (compte tenu des réserves exprimées au § 3.1.1), leur interprétation ne pourrait se faire que dans le cadre de la séquence de mesure complète. On est amené à se demander dans quelle mesure cela est représentatif des scènes naturelles et si le responsable de la conception d'un codec aurait la possibilité d'optimiser les caractéristiques de l'appareil pour les adapter à celles d'images d'essai connues.

3.2.2 Méthodes basées sur des images naturelles et des erreurs de codage

Il est important d'utiliser des séquences d'images naturelles dans toute évaluation d'un codec à mots de longueur variable. Si l'on tient compte de la faculté qu'ils ont de répartir l'utilisation des bits de codage sur toute l'image, on doit considérer de près la nature de chacune des parties de la séquence pour évaluer son caractère plus ou moins critique (voir le § 9). Il est recommandé que toutes les mesures objectives soient basées sur les erreurs de codage données par un codec recevant à son entrée une série d'images naturelles. La méthode de l'erreur quadratique moyenne examinée au § 3.1.3 peut aussi être utilisée pour les erreurs de codage dans les codecs à longueur de mot variable, mais l'interprétation des résultats est réservée aux spécialistes et même ainsi, seulement en complément d'évaluations subjectives. De même, la comparaison objective entre codecs au moyen de la technique EQMN ne doit être faite que par des experts en conception de codecs et uniquement lorsque les techniques à comparer sont très peu différentes (c'est-à-dire constituent des variantes d'un même algorithme) et que l'on peut admettre l'identité des processus de dégradation.

3.2.3 Méthodes basées sur des modèles de vision

Le principal inconvénient des mesures basées sur la méthode EQMN est qu'elle ne reconnaît pas la forte influence que la nature de l'image exerce sur la perception des dégradations par l'observateur. Comme signalé au § 3.1.4, l'optimisation d'un principe de codec nécessite de recourir à des expériences subjectives et à des données psycho-visuelles pour adapter la tolérance aux distorsions de l'observateur humain aux caractéristiques locales des régions de l'image. De cette manière, lorsqu'un codec à longueur de mot variable répartit dans toute l'image la capacité binaire de codage (et, partant, l'importance des erreurs de codage), cette opération est également adaptée aux caractéristiques visuelles. Toute méthode de mesure objective doit donc tenir compte des

propriétés du système visuel humain si l'on veut qu'elle donne des résultats en bonne corrélation avec les notes d'évaluation subjective. Le modèle de vision a pour fonction d'interpréter les erreurs de codage en tenant compte de l'image initiale dans laquelle elles se produisent.

Dans le texte ci-après, on part du principe qu'on ne peut avoir accès aux mécanismes internes des codecs. Si l'on peut obtenir des renseignements sur les modes d'adaptation, les performances des codecs à longueur de mot variable peuvent aussi être évaluées au moyen de la méthode des facteurs de dégradation (§ 3.1.1), comme cela est indiqué dans [Inoue et Hishiyama, 1984].

La création d'un modèle de vision doit faire intervenir deux niveaux de connaissance. Le premier concerne la visibilité d'une dégradation quelconque selon sa localisation dans l'image et le second porte sur la manière de déterminer comment la visibilité de toutes les dégradations doit être combinée pour conduire à une note de qualité globale. Il suffit donc de s'intéresser aux modèles prenant en compte les caractéristiques de dégradation des méthodes de codage numérique; il est par exemple, inutile de considérer les distorsions de nature géométrique ou sémantique. Les modélisations de la réponse du système visuel humain aux distorsions dues à la transmission de l'image se sont concentrées, sur les phénomènes au seuil de visibilité, ou au voisinage, ce qui convient pour la télévision de haute qualité (voir, par exemple, [Sakrison, 1977]). On sait peu de choses sur la modélisation de la réponse à des distorsions plus importantes.

Une étude importante sur la conception d'un modèle de la vue pour la prévision de la qualité d'image a été réalisée par [Lukas et Budrikis, 1982]. Leur article étudie la création de ce modèle et ses performances en tant qu'outil de prévision de la qualité subjective à partir d'un estimateur basé sur des mesures brutes d'erreur, l'un représentant le filtrage visuel non linéaire, et l'autre tenant compte des propriétés de masquage spatial et temporel de la vision. On a employé comme support de cette étude les processus de distorsion de la quantification linéaire, du codage MICD, du bruit gaussien additif et du filtrage passe-bas. Il est particulièrement important de noter pour la création d'un modèle de mesure de la qualité d'une séquence d'images que les observateurs ont tendance à noter l'image selon la distorsion présente dans sa région la plus dégradée et non selon la moyenne pour toutes celles-ci. Un autre modèle de vision a été établi plus récemment [Girod, 1988] et [Zetzsche & Hauske, 1989] pour les applications de codage numérique de l'image.

L'utilisation de modèles de vision est très prometteuse pour la détermination objective de la qualité d'image en présence non seulement de dégradations dues au codage numérique, mais aussi de défauts causés par d'autres processus non linéaires ou adaptatifs. On ne lui a malheureusement accordé que peu d'attention et les spécialistes sont incités à soumettre de nouvelles contributions à son sujet.

4. Mesure objective de la qualité d'image des codecs en présence d'erreurs de transmission

Dans un système de transmission réel, la liaison entre le codeur et le décodeur sera soumise à des influences qui perturberont les données acheminées, de telle sorte qu'une importante caractéristique d'un décodeur est sa réponse à la présence de ces erreurs de transmission. Dans un codec bien conçu, la réponse en question prendra la forme de distorsions transitoires locales de l'image décodée et le nombre de ces transitoires sera en relation avec la répartition statistique des erreurs dans la voie, tandis que leur nature dépendra de l'algorithme de codage employé, ainsi que du caractère plus au moins critique de

la séquence d'images présentée. Le but des expériences faisant intervenir les erreurs de transmission est généralement d'établir pour un codec une représentation graphique des dégradations perçues par l'observateur pour une série de taux d'erreur.

Un décodeur possède plusieurs niveaux de traitement qui peuvent déterminer sa réponse aux erreurs de transmission; certains peuvent être analysés mathématiquement ou être simulés sur ordinateur, tandis que d'autres exigent soit un certain degré d'évaluation subjective, soit la création d'un modèle de la réponse de l'observateur aux distorsions transitoires.

La première étape d'une analyse objective est de décrire aussi exactement que possible la manière dont les erreurs surviennent dans une liaison réelle, ce qui se fait généralement sous la forme d'un modèle statistique. Sous sa forme la plus simple, ce modèle admet que les erreurs se produisent aléatoirement et de manière indépendante (répartition de Poisson), mais l'on sait depuis longtemps, par observation, qu'en réalité les erreurs apparaissent en groupes ou en salves. On a proposé plusieurs modèles pour rendre compte de ce comportement, le plus connu étant basé sur une répartition de Neyman du type A (voir, par exemple, [Jones & Pullum, 1981]). Alors que la loi de Poisson est entièrement définie par un seul paramètre, le taux d'erreur binaire moyen, le modèle Newman A, en exige deux autres décrivant le degré de regroupement et la densité des erreurs dans chaque groupe. On ne dispose pas encore d'une recommandation permettant de fixer ces deux paramètres à des valeurs réalistes.

Conscients de la caractéristique en salves des erreurs de transmission, les concepteurs de codecs prévoient souvent une opération de réarrangement temporel des bits à transmettre, avant leur application à la voie. On obtient ainsi une dispersion des salves d'erreurs dans la voie lorsque les bits subissent le réarrangement inverse dans le décodeur, ce qui permet de mieux traiter les erreurs en question dans le système de correction qui suit. Celui-ci sera capable de corriger complètement un certain nombre d'erreurs en utilisant des données de redondances transmises dans une partie de la capacité de la voie, mais il subsistera quelques erreurs "résiduelles" qui parviendront au système de décodage de l'image. La répartition des erreurs résiduelles peut être calculée pour un codec et un modèle de voie donnés, mais il reste à étudier leur effet sur l'image décodée.

[CCIR, 1986-90b] suggère que les performances d'un codec donné vis-à-vis des erreurs de transmission soient jugées en deux parties, d'abord subjectivement pour déterminer la dégradation due à sa caractéristique de distorsion, puis objectivement, en tenant compte du taux d'erreur résiduel obtenu par le calcul à partir des considérations qui précèdent. On ne dispose pas actuellement de résultats expérimentaux confirmant ce principe, mais il pourrait toutefois constituer la première étape d'une mesure entièrement objective si la réponse de l'observateur aux différentes transitoires du codec peut être chiffrée. Il est important de noter que certains des bits transmis sont plus sensibles que d'autres aux perturbations, ce qui signifie que la réponse du codec à une erreur résiduelle sur un seul bit peut beaucoup varier et peut aussi dépendre du caractère plus ou moins critique de la séquence d'images à la source. C'est ainsi que dans les codecs intertrames, les transitoires provenant d'erreurs résiduelles peuvent se cantonner dans les parties fixes de l'image jusqu'au moment où des mesures sont prises pour les éliminer par rafraîchissement. Enfin, l'une des particularités de certains codecs à mots de longueur variable est qu'ils peuvent détecter quelques violations du codage causées par les erreurs de transmission et utiliser cette connaissance pour essayer de dissimuler les distorsions transitoires. La dissimulation ne réussit pas avec toutes les erreurs, mais elle améliore généralement la qualité subjective de l'image produite, ce dont il faut tenir compte dans l'évaluation objective d'un codec.

5. Evaluation subjective de la qualité d'image des codecs

Des progrès sont en cours, mais l'on ne dispose pas actuellement d'une expérience suffisante pour pouvoir donner des précisions sur les méthodes d'évaluation objective de la qualité d'image des codecs. En matière d'évaluations subjectives pour lesquelles on dispose d'une expérience considérable, on peut faire des recommandations sur les conditions expérimentales et sur la méthodologie. Il ne faut cependant pas oublier en fixant des objectifs de qualité ou de dégradation que les méthodes actuelles ne peuvent pas donner des notes subjectives absolues, mais au contraire, des résultats qui sont partiellement affectés par les conditions de référence au d'ancrage. On peut adopter les mêmes méthodologies pour les codecs à longueur de mot fixe ou variable, inter ou intratrames, mais leur nature pourra influencer le choix des séquences d'essai (voir le § 9).

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

5.1 Evaluation de la qualité intrinsèque

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

5.1.1 Conditions d'observation et choix des observateurs

Il est conseillé de respecter dans ce domaine les prescriptions figurant au § 2.4 de la Recommandation 500 pour la télévision de définition conventionnelle et dans la Recommandation 710 pour les codecs de télévision à haute définition.

5.1.2 Séquences d'images d'essai

Il est recommandé d'utiliser dans les essais au moins six séquences d'images, plus une supplémentaire destinée à une démonstration avant le début de l'expérience. La durée d'une séquence doit être d'environ 10 s, mais il convient de noter que les observateurs préféreront peut-être une durée de 15 à 30 s [Inoue, 1988] [CCIR, 1986-90c]. Cette séquence doit être comprise entre "moyennement critique" et "critique" pour l'application à débit binaire réduit considérée (voir le § 9).

5.1.3 Méthodologie des essais

Lorsque la plage de qualité à étudier est étroite, comme ce sera normalement le cas pour les codecs de télévision, la méthodologie à utiliser est celle du double stimulus avec échelle de qualité continue décrite dans la Recommandation 500. La séquence de source initiale est utilisée comme référence. Des conversations se poursuivent sur la durée des séquences de présentation au sein du GTI 11/4 [CCIR, 1986-90d, e]. Dans de récentes expériences du GTI 11/7 sur des codecs destinés à la vidéo en composantes 4:2:2

[CCIR, 1986-90a, f] dont les résultats sont donnés dans [CCIR, 1986-90g], il a été jugé avantageux de modifier la présentation par rapport à celle décrite dans la Recommandation 500-3 (MOD F). Des images composites ont été utilisées à titre de référence supplémentaire pour donner un niveau de qualité plus faible par rapport à celui auquel était estimée la qualité du codec.

5.2 Evaluation de la qualité après post-traitement

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

5.2.1 Conditions d'observation et choix des observateurs

Voir le § 5.1.1.

5.2.2 Séquences d'images d'essai

Du fait de contraintes matérielles ou de la nécessité d'évaluer un codec pour plusieurs opérations de post-traitement, le nombre des séquences d'images utilisées peut être au minimum de trois, avec une supplémentaire pour la démonstration. La nature des séquences dépendra du type de post-traitement étudié, mais elles devront être comprises entre "modérément critique" et "critique" pour la télévision à débit binaire réduit et pour l'opération en question. La durée des séquences devra être d'environ 10 s, mais il convient de noter que les observateurs préféreront peut-être une durée de 15 à 30 s [Inoue, 1988] [CCIR, 1986-90c].

5.2.3 Méthodologie des essais

la méthodologie à utiliser est celle du double stimulus à échelle de qualité continue décrite dans la Recommandation 500. Dans ce cas, la référence sera cependant la séquence à la source soumise au même post-traitement que les images décodées. Si l'on juge avantageux de prévoir une référence de plus basse qualité, elle doit aussi être soumise au même post-traitement. Pour les essais décrits dans [CCIR, 1986-90f], on a légèrement modifié la présentation décrite dans la Recommandation 500.

6. Evaluation subjective de la dégradation de la qualité d'image des codecs du fait des erreurs de transmission

On considère au § 4 la manière dont les erreurs de transmission sont traitées par un décodeur numérique en vue d'envisager comment on pourrait aborder l'analyse objective de la qualité d'image. Il pourrait être utile d'évaluer subjectivement les dégradations en fonction du taux d'erreur sur la liaison entre le codeur et le décodeur. On ne possède actuellement qu'une connaissance expérimentale des statistiques d'erreurs réelles insuffisante pour pouvoir faire des recommandations sur les paramètres d'un modèle rendant compte du groupement des erreurs en salves. Tant que l'on ne disposera pas de données suffisantes, la loi de Poisson pourra être appliquée aux erreurs. On trouvera dans [CCIR, 1986-90f] quelques précisions sur les facteurs de corruption pour les niveaux hiérarchiques de transmission de 34, 45 et 140 Mbit/s.

6.1 Séquences d'images d'essai

Comme l'on doit étudier la qualité du codec pour une série de taux d'erreur, les contraintes matérielles conduisent à suggérer que trois séquences, plus une pour la démonstration, seraient peut-être suffisantes. La durée d'une séquence doit être d'environ 10 s, mais il convient de noter que les observateurs préféreront peut-être une durée de 15 à 30 s [Inoue, 1988] [CCIR, 1986-90c]. Cette séquence doit être comprise entre "moyennement critique" et "critique" pour la télévision à débit binaire réduit (voir le § 9).

6.2 Choix des taux d'erreur

On doit choisir un minimum de cinq taux d'erreur, mais de préférence davantage, répartis logarithmiquement et couvrant toute la gamme pour laquelle le codec donne des dégradations allant de "imperceptible" à "très gênant".

6.3 Méthodologie des essais

Comme les essais couvriront toute la plage des dégradations, l'échelle de dégradation à double stimulus (UER) est applicable et doit être utilisée. Elle est décrite dans la Recommandation 500.

6.4 Note sur l'utilisation de très faibles taux d'erreur

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

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

7. Comparaisons subjectives entre codecs

Lorsqu'il est inutile de juger la qualité ou la dégradation absolue d'un codec et qu'on ne demande que de classer les systèmes ou lorsqu'on souhaite confirmer un classement obtenu par la méthode du double stimulus, on doit utiliser une technique de comparaison entre couples de stimuli. Ce point figure au § 4.2 de la Recommandation 500.

Telle qu'elle est décrite, la méthode permet des comparaisons précises et donne le moyen de déterminer la relation entre couples de systèmes. [CCIR, 1986-90h] décrit une extension de cette méthode au classement de la qualité ou de la dégradation pour plus de deux systèmes. Selon ce principe, on établit un classement global à partir de ceux donnés par les observateurs pour tous les couples possibles de séquences d'images.

L'analyse est compliquée par le fait qu'un observateur peut juger que l'image A est meilleure que l'image B, elle-même meilleure que l'image C, tout en préférant C à A. On parle alors de "triade intransitive". Le traitement statistique de la transitivité pour chaque observateur est étudié dans [CCIR, 1986-90h, i] car l'existence d'un accord statistiquement significatif entre observateurs est un facteur important.

Un problème posé par la méthode est que le nombre de présentations nécessaires augmente comme le carré du nombre de séquences d'essai et de codecs, ce qui peut le faire devenir impossible à réaliser.

8. Distorsions dans la transmission mixte analogique-numérique

Jusqu'à présent, on a étudié séparément les problèmes de définition de la qualité d'image pour les systèmes analogiques et numériques. Si le principe de l'indépendance psychologique des phénomènes de dégradation de la qualité d'image mentionné au § 3.1 est admissible, la technique décrite dans ce paragraphe est aussi applicable aux systèmes mixtes. En d'autres termes, les dégradations peuvent être réparties dans l'une des catégories suivantes, du point de vue de l'indépendance psychologique:

- a) dégradations causées uniquement par la section analogique;
- b) dégradations causées uniquement par la section numérique;
- c) dégradations causées par les sections analogique et numérique, susceptibles d'être des facteurs indépendants dans chaque système.

Les dégradations des groupes a) et b) seront traitées comme des facteurs indépendants et l'on a déjà proposé au CCIR une fonction destinée à estimer la qualité d'image globale dans un tel cas. Cette fonction d'estimation est applicable lorsqu'il existe simultanément plusieurs facteurs psychologiquement indépendants.

En revanche, dans le groupe c), où les dégradations provenant des deux sections sont si semblables qu'on ne peut pas les considérer comme indépendantes, il faudra trouver un nouveau moyen d'attribuer les dégradations aux sections analogique et numérique avant d'appliquer la fonction d'estimation mentionnée plus haut.

[Inoue, 1987] donne un exemple de résultats d'études dans un cas de ce genre. L'article examine la combinaison d'un bruit aléatoire dans le système analogique et d'un bruit granulaire dans un système de codage MICD intratrame à mots de longueur invariable pour démontrer qu'il est possible de remplacer une mesure physique dans le système analogique par une valeur corrigée basée sur les différences de sensibilité de l'oeil.

9. Choix des images d'essai pour l'évaluation des codecs numériques

On insiste dans tout le présent document sur l'importance d'essayer les codecs au moyen de séquences d'images critiques pour la réduction de débit binaire en télévision. On peut donc raisonnablement se demander qu'elle est la criticité d'une séquence donnée pour une application déterminée de la réduction du débit binaire ou se préoccuper de savoir si une séquence est plus critique qu'une autre. Une réponse simple, mais pas particulièrement utile, est d'affirmer que la notion de criticité a une signification très différente selon les codecs. C'est ainsi qu'une image fixe très détaillée pourrait être critique pour un codec intratrame, tandis que pour un système intertrames capable d'exploiter les similitudes entre images, la même scène ne poserait aucun problème. Certaines séquences où figurent des textures mouvantes et des mouvements complexes seront critiques pour toutes les catégories de codecs et ce sont donc elles qu'il est plus utile de produire ou de reconnaître. Les mouvements complexes peuvent être d'une forme prévisible par l'observateur, mais non par les algorithmes de codage; c'est le cas des déplacements tortueux périodiques.

On a découvert une grandeur simple mais utile basée sur la mesure de l'entropie adaptative intertrame-intratrame en essayant de chiffrer statistiquement la criticité [CCIR, 1986-90j]. Cette méthode a été utilisée pour "calibrer" les séquences d'images que le GTI 11/7 propose d'employer dans les essais de codecs à 34, 45 et 140 Mbit/s et s'est révélée utile dans le choix de celles à adopter. La manière la plus facile de réaliser ces mesures sur les séquences est de transférer celles-ci à des ordinateurs de traitement de l'image, puis en les soumettant à un logiciel d'analyse.

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

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

Codecs intertrames à mots de longueur fixe: les scènes choisies doivent toutes comporter des mouvements et le plus possible des détails suivants: régions texturées mobiles (certaines en couleur), objets comportant des contours fins et fortement contrastés (certains colorés) se déplaçant dans diverses orientations perpendiculaires aux arêtes. Une au moins doit comporter un bruit à la source à peine perceptible et une autre (au minimum) doit être artificielle.

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

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

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

Le GTI 11/7 a rassemblé une collection complète de séquences d'essai enregistrées au format D1 en composantes 4:2:2. Le Rapport 1213 donne des précisions à leur sujet et indique les critères (qui peuvent s'appliquer à d'autres normes de représentation d'image) sur la base desquels elles sont préparées.

REFERENCES BIBLIOGRAPHIQUES

- GIROD, B. [1988] - A Model of Human Visual Perception for the Reduction in Redundancy of Television Luminance Signals, PhD. Thesis University of Hannover 1988 (en allemand).
- HISHIYAMA, K. et INOUE, M. [septembre, 1984] - Physical Measures for Interframe Coded Picture Quality, Review of ECL, Vol 32, N° 5.
- INOUE, M. [janvier, 1988] - The Influence of Picture Presentation Period on Subjective Evaluation, Tech. Rep. of IEICE (Japon), IE 87-105 (en japonais).
- INOUE, M. [mai, 1987] - The Proposed Method for Noise Specifications to Mixed Analogue-Digital Video Transmission Systems, JITEJ, Vol 41, N° 5 (en japonais).
- INOUE, M. et HISHIYAMA, K. [1984] - Trade-off between Information Suppression Effect and Picture Quality Degradation with Interframe Coding, Review of the Electrical Communication Laboratories (Japon), Vol. 32, N° 5.

- JONES, W.J. et PULLUM, G.G. [mars, 1981] - Error Performance Objectives for Integrated Services Digital Networks Exhibiting Error Clustering, Proc. IEE Conf. on Telecom Transmission, Londres, Publ. N° 193.
- KOBAYASHI, Y. [mai, 1977] - Picture Quality Evaluation Method for Digitally Encoded Video Signals, Trans. Inst. Elect. Com. Engrs. Japon, Vol J60-B, N° 5 (en japonais).
- LUKAS, J.X.J. et BUDRIKIS, Z.L. [juillet, 1982] - Picture Quality Prediction Based on a Visual Model, IEEE Trans. Com., Vol COM-30, N° 7.
- ROUFS, J., DE RIDDER, H. et WESTERINK, J. [1989] - Perpetual Image Quality Metrics, Institute for Perception Research, Eindhoven, Manuscript MS 692.
- SAKRISON, D.J. [novembre, 1977] - On the Role of the Observer and a Distortion Measure in Image Transmission, IEEE Trans. Com., Vol COM-25, N° 11.
- ZETZSCHE, C. et HAUSKE, G. [juillet, 1989] - Principal Features of Human Vision in the Context of Image Quality Models, Proc. IEE Int. Conf. on Image Processing and Applications, Warwick, Publ. N° 307.

Documents du CCIR

[1982-86]: 11/26 (CMTT/31) (Japon)

[1986-90]: a. 11/498 (GTI 11/7); b. GTI 11/7-192 (Japon); c. 11/422 (Japon); d. GTI 11/4-154 (Japon); e. GTI 11/4-158 (France); f. GTI 11/4-182(Rév.2) (GTI 11/7); g. 11/498 (GTI 11/7); h. GTI 11/4-175 (Allemagne (Rép. féd. d')); i. GTI 11/4-156 (Allemagne (Rép. féd. d')); j. GTI 11/7-261 (Royaume-Uni).
