RAPPORT 1082-1

ETUDES EN VUE DE L'UNIFORMISATION DES METHODES D'EVALUATION DE L'IMAGE

(Question 3/11, Programme d'études 3A/11)

(1986-1990)

1. Introduction

La Recommandation 500 fait l'objet de révisions périodiques afin d'indiquer quelles sont les méthodes disponibles qui semblent les meilleures pour évaluer la qualité des images en laboratoire dans des conditions bien définies. Il faut revoir périodiquement ces méthodes pour suivre l'évolution des études de nouveaux systèmes et de la méthodologie elle-même.

Bien que les méthodes décrites dans les § 2 et 3 de la Recommandation 500 aient été étudiées et conçues avec soin, en fonction des connaissances du moment, elles présentent certains défauts. S'il en existe d'autres qui en sont exemptes, elles ont vocation à remplacer les méthodes actuelles.

Les principaux défauts des méthodes décrites actuellement dans les § 2 et 3 sont les suivants:

- Les différences de sens entre les descripteurs des divers échelons de la notation ne sont pas forcément uniformes. On sait qu'elles varient sensiblement avec les groupes linguistiques, culturels, et d'un individu à l'autre. Pour dégager des résultats, on suppose que ces différences sont uniformes; c'est une approximation, de même par conséquent que l'interprétation des résultats en vue d'en tirer une mesure cohérente de la qualité ou de la dégradation absolue. En fait, l'erreur qu'on risque de commettre quand on estime les différences à partir des résultats peut atteindre $\pm$ 50%.

- En raison notamment des acceptations diverses des descripteurs mentionnés ci-dessus, on considère que la corrélation entre les conclusions obtenues dans divers laboratoires n'est pas suffisamment bonne pour comparer des résultats absolus et des systèmes différents qui présentent de faibles défauts ou une bonne qualité. Toutefois, les classements sont cohérents.

- La stabilité des méthodes des § 2 et 3 de la Recommandation 500 résulte en partie du recours systématique à une référence de haute qualité. Mais il arrive qu'on ne dispose pas d'une référence de haute qualité; en ce cas, ces méthodes ne s'appliquent pas.

- Les méthodes à double stimulus prennent plus de deux fois plus de temps que les méthodes à simple stimulus et reviennent donc plus cher.

Le présent Rapport décrit des études de développement de nouvelles méthodes censées fournir davantage d'information et éviter les défauts signalés ci-dessus. Les études ont porté sur les points suivants:

- échelle de rapports (estimation de la qualité en grandeur numérique);

- échelles graphiques (estimation des différences de sens entre les descripteurs);

- échelles catégorielles numériques;

- échelles multidimensionnelles;

- comparaison deux à deux;

- mesure du seuil de visibilité.

Seules les méthodes qui sont tout à fait mises au point et qui présentent de nets avantages par rapport aux méthodes actuellement recommandées peuvent prétendre figurer dans la Recommandation 500.

Le présent Rapport décrit aussi des travaux effectués récemment pour savoir si on pourrait, au moyen de l'échelle graphique, estimer des dégradations comme celles qui sont dues au bruit.

2. Echelles de rapports

2.1 Introduction

Etant donné qu'elle se prête à la plus large gamme et à la plus grande variété d'opérations statistiques, l'échelle de rapports permet à l'expérimentateur non seulement de classer les éléments à échelonner, mais aussi de décrire les magnitudes relatives d'un des attributs de ces éléments. On peut déterminer, par exemple, non seulement que l'image A est meilleure que l'image B mais aussi combien de fois meilleure. Une échelle catégorielle ne permet que d'établir un classement, pas de définir les intervalles entre échelons. L'estimation en grandeur est la méthode psychophysique la mieux appropriée pour établir des échelles de rapport de la qualité de l'image.

La méthode d'estimation magnitudes est utilisée dans le monde entier depuis le milieu des années 50. Les échelles se traduisent par des rapports, donc par des données séparées par des intervalles égaux. A partir de ces données, il est légitime de calculer des moyennes géométriques ou harmoniques aussi bien qu'arithmétiques, et de calculer des variations de pourcentages aussi bien que des écarts types. La précision ainsi améliorée permet de décrire plus complètement les données. Les observateurs construisent eux-mêmes leurs propres échelles; ils évitent de ce fait un grave défaut des échelles fixes, à savoir l'attribution éventuellement inappropriée de désignations verbales et des échelles numériques.

2.2 Méthodologie des expériences

2.2.1 Procédure

La méthode d'estimation des magnitudes permet aux observateurs de créer leurs propres échelles au fur et à mesure du déroulement de l'expérience. On lui présente une série d'images dans un ordre quelconque et on lui demande d'attribuer une valeur numérique à la qualité de chacune d'elles. Il faut que les instructions contiennent les renseignements suivants et soient libellées selon le modèle ci-après:

On va vous présenter une série d'images dans un ordre quelconque. Vous avez à apprécier la qualité de l'image en attribuant un numéro à chacune d'elles. Donnez à la première image un numéro qui vous semble lui convenir. Puis attribuez des nombres proportionnés aux présentations suivantes en les comparant pour exprimer votre impression subjective. Il n'y a pas de limite à l'intervalle des nombres que vous pouvez utiliser. Il peut s'agir de nombres entiers, décimaux ou fractionnaires. Essayez de faire en sorte que chaque nombre corresponde à la qualité de l'image telle que vous la percevez. Si, par exemple, une image vous paraît trois fois meilleure qu'une autre, attribuez-lui un nombre trois fois plus grand; si elle vous semble cinq fois moins bonne, attribuez-lui un nombre cinq fois plus petit.

La gamme et le nombre des stimuli devraient être suffisamment importants pour cette expérience particulière, cela afin de ne pas limiter l'observateur à un petit nombre de conditions mais au contraire de lui permettre d'utiliser tous les critères disponibles pour procéder à ses évaluations de qualité. Chaque stimulus est d'ordinaire présenté deux fois; on a constaté que le fait de présenter un stimulus plus de deux fois ne donne pratiquement aucune information supplémentaire.

2.2.2 Appréciation de l'"idéal"

Afin d'établir une référence pour pouvoir comparer les résultats d'essai obtenus dans divers laboratoires avec des systèmes de télévision ou des images d'essais différents, etc., il faut demander aux observateurs, à la fin de chaque série d'essais, d'attribuer une valeur numérique à la qualité d'image qu'ils considèrent "idéale". "Idéal" est censé se référer à la meilleure qualité d'image possible imaginable produite par n'importe quel système de formation d'images. Lorsqu'on analysera les données (voir § 2.2.6) l'appréciation "idéal" (c'est-à-dire le nombre en question) sera normalisé pour prendre la valeur 100 et constituer une norme uniforme.

2.2.3 Participants

On a constaté que les rapports des moyennes géométriques se stabilisent à partir de 15 observateurs (évidemment, il peut au besoin y en avoir davantage).

2.2.4 Images d'essai

Les images d'essai qui conviennent le mieux dépendent de la nature de l'expérience.

2.2.5 Présentation

La séquence des images doit être aléatoire, étant entendu que la même image (scène ou séquence d'essai) ne devrait pas être présentée deux fois de suite au même niveau de qualité. Il convient, si possible, de présenter une séquence aléatoire différente à chaque observateur. Il faut faire varier le stimulus initial pour chaque observateur, mais il n'est pas nécessaire d'en faire de même pour le niveau de qualité. Il est conseillé de commencer chaque séance quelque part vers le milieu de la gamme mais non à l'une de ses extrémités.

Une séance d'observation doit durer environ une demi-heure, explications et manœuvres préliminaires comprises. La séquence d'essai pourrait commencer par quelques images typiques de la gamme des qualités d'image (mais il ne faut absolument pas dire aux observateurs quelle sera cette gamme). Les jugements portés sur ces présentations préliminaires n'entrent pas en ligne de compte dans les résultats finals.

2.2.6 Normalisation et moyenne des estimations de magnitude

Lorsqu'on dispose d'un ensemble d'estimations de magnitude, c'est leur moyenne géométrique qui constitue la mesure exacte, d'ailleurs la plus communément utilisée, de leur tendance centrale. Elle tient compte de la distribution des réponses et présente cet avantage qu'elle empêche un jugement extrême d'influencer exagérément le résultat. Elle donne une estimation impartiale de la valeur attendue des logarithmes des estimations quantitatives. Bien que les différents observateurs aient pu attribuer des notes différentes au premier stimulus, il est inutile de procéder à une normalisation avant de calculer les moyennes. Les rapports des moyennes géométriques ne sont pas affectés, même si les observateurs utilisent des unités différentes pour leurs échelles subjectives. Cependant, la normalisation deviendra nécessaire pour certaines opérations statistiques ultérieures et pour les comparaisons entre laboratoires. A cet effet, pour chaque résultat fourni par un observateur, il faudra effectuer les calculs ci-après afin d'effectuer une normalisation par rapport à l'«idéal».

On normalisera les moyennes géométriques de manière à attribuer à l'«idéal» la valeur normalisée 100. Pour ce faire, on calculera la moyenne géométrique des réponses numériques. On multipliera ensuite les moyennes géométriques par le facteur commun $100/R_i$ (R_i étant la valeur numérique de la réponse «idéale»). Moyennant cette simple opération, l'«idéal» est défini par 100 et, du même coup, les réponses moyennes sont ajustées à tous les autres stimuli dans la même proportion.

2.3 Etudes sur la performance de l'échelle des rapports

2.3.1 Utilisation d'une échelle de notation catégorielle à stimulus unique et d'une échelle de rapports: étude comparative

2.3.1.1 Introduction

On a procédé à deux séries d'expériences sur la qualité de l'image [CCIR, 1982-86a] au moyen de l'échelle de notation CCIR de la qualité et, à titre de comparaison, de la méthode d'estimation des magnitudes. L'objectif était d'étudier l'influence du contexte sur les deux méthodes d'essai. L'intérêt pour ces questions a été renouvelé par l'apparition d'images de télévision très améliorées et par l'extension de la gamme de la qualité des images qui en est résultée.

2.3.1.2 Procédures

Méthode d'essai

On testait individuellement chaque observateur au moyen de deux méthodes d'estimation de la qualité. Pour évaluer l'échelle de notation catégorielle, on se servait de l'échelle de qualité du CCIR comme méthode d'estimation de la qualité des images d'essai. Les notes moyennes étaient calculées pour chaque image d'essai. L'évaluation de l'estimation en grandeur se faisait selon les méthodes décrites dans le § 2.2 du présent Rapport.

Appareils utilisés et mode opératoire

Les observateurs étaient assis à une distance d'observation égale à trois fois la hauteur de l'image. Dans les deux cas, les images étaient présentées sur un écran de 19 pouces (48,3 cm) de diagonale (tube à rayons cathodiques), fonctionnant avec une fréquence de trame de 60 Hz.

Il y avait quatre niveaux de qualité de l'image dans l'un des essais et cinq dans l'autre: A pour le NTSC à 525 lignes avec décodeur à filtre à coupure brusque, B pour le NTSC à 525 lignes avec décodeur à filtre en peigne, C pour les signaux RGB provenant directement de la caméra et D pour la haute définition. Toutes ces images étaient considérées comme de bonne ou d'excellente qualité (gamme étroite). Dans l'essai avec gamme étendue, les niveaux de stimulus X et Y étaient du NTSC à 525 lignes, avec filtre à coupure brusque et adjonction de bruit; les rapports signal/bruit étaient, dans ce cas, de 22 dB pour X et de 32 dB pour Y. Cet essai comportait moins de données RGB car chaque observateur n'évalue les RVB qu'une fois, en tant que dernière appréciation de chaque séance d'essais.

Environ cinq semaines s'écoulaient entre les deux essais.

Participants

Les 67 observateurs avaient une acuité visuelle et une vision des couleurs normales ou corrigées. Aucun d'eux n'avait jamais pris part à des expériences d'estimation en grandeur mais certains connaissaient déjà les échelles à 5 notes.

Les participants se répartissaient en trois groupes:

Un groupe de 9 «experts» avait été pris parmi le personnel du laboratoire. Il s'agissait d'hommes, âgés de 27 à 65 ans, faisant partie des techniciens de la télévision.

Un autre groupe, de 47 personnes, avait été constitué à l'aide d'agents non techniciens; c'était des hommes et des femmes âgés de 33 à 60 ans.

Le troisième groupe se composait de 11 lycéens, garçons et filles, de 16 ans. Avec eux, on atteint le total précité de 67 observateurs.

Conditions d'observation

On se conformait en général aux conditions que prescrit la Recommandation 500, à ceci près que la distance d'observation était égale à trois fois la hauteur de l'image.

Images d'essai

Les trois images d'essai étaient des diapositives en couleur de 8 × 10 pouces (8 × 25,4 cm) de la région de Stamford, Connecticut. Elles étaient éclairées par un Porta-Pattern. On les avait choisies exprès en vue d'obtenir la meilleure qualité possible ainsi qu'un niveau raisonnable d'information de fréquence spatiale élevée.

Deux caméras fonctionnaient sur le Porta-Pattern: l'une de qualité moyenne à 525 lignes, l'autre pour TVHD à 1125 lignes. Les images étaient transmises des caméras aux codeurs ou, selon le cas, directement sur les écrans.

2.3.1.3 Résultats

Essais avec échelle à 5 notes

———————————————————————————————. Les trois niveaux de qualité qui étaient communs aux deux essais divergent quelque peu. Ce sont les images avec filtre à coupure brusque qui présentent le plus grand décalage (de 1,88 à 3,56), soit 63% de la gamme totale entre la réponse minimale et la réponse maximale. En d'autres termes, le décalage des notes, lorsqu'on passait de l'essai à gamme étroite à l'essai à gamme étendue, était presque aussi grand que la gamme totale des notes attribuées lors de l'essai à gamme étroite.

On notera aussi à cette occasion que les qualificatifs sont apparemment dénués de signification: une image estimée «médiocre» était devenue «bonne».

Essais d'estimation des magnitudes

On rencontre ici des variations analogues, mais elles sont bien plus petites qu'avec l'échelle à cinq notes. Les plus fortes ont été de nouveau observées avec le filtre à coupure raide mais elles n'atteignaient alors que 34% (au lieu de 63%).

On notera que lorsque la gamme des stimuli s'élargit, les observateurs produisent des nombres plus dispersés. Pour l'essai à gamme étroite, les nombres s'étendent sur un intervalle de 43 (19,5 à 62,5) alors que pour l'essai à gamme large l'intervalle est presque égal à 60 (4,12 à 64,5). C'est une preuve supplémentaire que l'échelle de grandeur est plus appropriée et s'adapte plus naturellement. Les observateurs réagissent mieux à une gamme plus étendue de qualité du stimulus, en élargissant leur échelle de notation. Avec l'échelle de qualité, les nombres et les désignations restent les mêmes lorsque la gamme des stimuli s'étend.

On notera enfin avec intérêt que les résultats obtenus avec les élèves, relativement plus spontanés, varient très peu avec la gamme des stimuli, c'est-à-dire qu'il y a peu de variations entre les essais à gamme étroite et à large gamme.

2.3.1.4 Conclusions

La modification de la gamme des stimuli affecte nettement moins les échelles de rapports que les échelles catégorielles à 5 notes.

Les échelles de rapports n'exigent pas d'interprétation linguistique. Il suffit que le participant ait une notion de la proportionnalité.

Les réponses numériques des échelles de rapports donnent des intervalles et des rapports significatifs. Elles indiquent donc aussi quantitativement la supériorité de qualité d'une image par rapport à une autre.

2.3.2 Usage comparé d'une échelle de qualité à double stimulus et d'une échelle de rapports

2.3.2.1 Introduction

On a procédé en France à deux autres séries d'expérience sur la qualité de l'image [CCIR, 1986-90 b]: au moyen de l'échelle de qualité à double stimulus et, à titre de comparaison, avec la méthode d'échelle de rapports. L'intérêt de ces expériences a déjà été exposé au § 2.3 mais, dans le cas présent, on a remplacé la simple méthode d'échelle avec notes de qualité par la méthode de notation continue de la qualité à double stimulus.

2.3.2.2 Procédure

Méthode d'essai

Ce sont les mêmes qu'au § 2.3.1, à ceci près que :

- les participants étaient groupés 4 par 4;
- l'ordre des présentations était le même pour tous les participants;

- la méthode à double stimulus était conforme à la seconde variante du § 3 de la Recommandation 500;

- avec les instructions étaient présentés quelques exemples d'image.

Mode opératoire et organisation

La distance d'observation des participants était égale à 6H. La fréquence trame était 50 Hz: la diagonale des écrans mesurait 51 cm.

Au moyen des divers codecs ci-après, on appréciait deux gammes de qualité (ou de dégradation):

i) gamme de dégradations étroite: RVB, MIC-D1, MIC-D2 et SECAM

ii) gamme de dégradations large: RVB, MIC-D1, SECAM, MIC-D1 + bruit et SECAM + bruit.

Participants

Il y avait au moins 15 participants pour chaque essai. Chacun d'eux ne participait qu'à un essai. Ce n'était pas des experts.

Conditions d'observation

Celles de la Recommandation 500.

Image d'essai

Quatre diapositives d'essai de l'UER.

2.3.2.3 Résultats

On a étudié quatre paramètres:

- Trois types de stabilité:

  - intragroupe: le même groupe a pris part deux fois à la même expérience;

  - intergroupe: on a comparé les résultats obtenus avec deux groupes différents;

  - influence du contexte (gamme): on a comparé les deux gammes.

- et sensibilité: comparaison du classement des faibles dégradations.

Pour vérifier que les différences étaient significatives, on s'est servi du test "t" de Student.

Sur toutes les expériences, on a effectué deux analyses pour avoir des résultats absolus et relatifs. En fait, on procède souvent à deux types d'évaluation: évaluation de la dégradation par rapport à une référence et évaluation de la qualité absolue. Ainsi, pour s'assurer dans les deux cas de la valeur de chacune des méthodes, on dispose du résultat du traitement des notations directes et de la différence entre la notation de la référence et celle de l'essai.

Les résultats des essais intragroupe ont présenté une stabilité acceptable; le "t" est resté dans l'intervalle de confiance à 90% (t = 1,7) et les écarts types sont semblables.

La possibilité de comparer les mesures d'un laboratoire à un autre dépend des résultats des essais intergroupe. Les deux procédures sont équivalentes car les valeurs de "t" sont proches de celle indiquée ci-dessus, ou inférieures. Il faudrait que les divers laboratoires poursuivent avec ces méthodes l'étude de la stabilité intergroupe des appréciations.

Les résultats de l'essai de comparaison gamme large-gamme étroite montrent que, dans le cas des résultats absolus, la valeur de "t" dépasse largement l'intervalle de confiance à 90%. Quant aux résultats relatifs, la méthode d'estimation en grandeur peut donner un faible "t", ce qui montre que cette évaluation est très stable. Cette différence entre les traitements absolus et relatifs montre que les résultats différents qui découlent de la modification de la gamme des dégradations se traduisent seulement par un glissement global dans le cas de la méthode d'estimation en grandeur.

On a finalement étudié la sensibilité des deux méthodes en comparant les classifications des dégradations dont l'importance est voisine dans chacune des expériences. Les deux méthodes semblent pouvoir chacune fournir un classement des faibles dégradations mais ce n'est pas le même, car il semble que les participants n'aient pas les mêmes critères pour les deux méthodes. La méthode du double stimulus semble inciter à un examen localisé dont le résultat est que le SECAM est le meilleur système. L'échelle de rapports semble conduire à une analyse globale qui favorise les dégradations numériques (MIC-D1 et MIC-D2).

2.3.2.4 Conclusion

De cet examen des méthodes on tire les conclusions suivantes:

- En pratique, la méthode de l'échelle des rapports modifiée convient à l'évaluation subjective actuelle des images de télévision.

- Aucune méthode ne peut donner sans référence des notes absolues fiables.

- Lorsque, pour chacun des essais, la gamme des dégradations est la même, la méthode du double stimulus aussi bien que celle de l'échelle des rapports sont assez stables pour permettre de comparer des résultats obtenus par différents laboratoires.

- Si la gamme des dégradations n'est pas la même, seule la méthode de l'échelle des rapports est assez stable pour que d'une expérience à l'autre les résultats relatifs soient comparables et encore à condition que les essais comprennent une référence implicite identique. Il faut donc une référence par type d'évaluation subjective: par exemple, télévision classique, télévision à haute définition.

- Les critères sur lesquels les participants fondent leur évaluation ne sont pas forcément les mêmes pour les deux méthodes. La méthode à échelle de rapports paraît mieux adaptée à une évaluation globale de la qualité subjective de l'image.

3. Echelle graphique

3.1 Introduction

Les échelles graphiques ont servi à déterminer les intervalles perçus entre dénominations descriptives. On a échelonné les adjectifs et les adverbes pour déterminer dans quelle mesure ils modifiaient les substantifs et les verbes. L'échelle de qualité de la Recommandation 500 du CCIR consiste en cinq termes qualitatifs qu'on a échelonnés pour déterminer l'amplitude des intervalles en allemand, anglais (Etats-Unis d'Amérique), français et italien. On a été surpris de voir comme ces intervalles étaient proches [Jones et McManus, 1986].

Les résultats des essais à échelle graphique sont valables par essence. Les participants énoncent leurs appréciations dans leur langue maternelle, sans contrainte et sans avoir à recourir à une interprétation numérique. On s'est servi de ces échelles pour estimer la qualité d'images en demandant aux participants d'indiquer sur une ligne le point qui correspond le mieux sur l'échelle à la qualité de l'image [Jones, 1986].

Le GTI 11/4 estime qu'en raison de son extrême simplicité et de la facilité de sa mise en oeuvre cette méthode d'évaluation subjective pourrait devenir une utile méthode d'essai internationale. On espère que d'autres administrations reprendront ces études dans leur langue propre, conformément aux instructions et directives suivantes.

3.2 Méthodologie expérimentale des essais

3.2.1 Procédure

3.2.1.1 Tracer sur des feuilles de longues lignes verticales (dans l'étude d'origine, le format des feuilles était 21 x 29,7 cm et la longueur des lignes 18 cm) et inscrire les dénominations extrêmes à chaque bout. Encadrer dans un coin un des mots de test (un seul par page).

3.2.1.2 Disposer les feuilles en autant de séquences aléatoires différentes que possible.

3.2.1.3 Remettre un jeu de feuilles à chaque participant. Demander au participant de tracer sur la ligne de chaque page une marque à l'endroit où, selon lui, le mot inscrit dans le coin se situe par rapport aux deux extrêmes. Traiter ainsi toutes les pages; ne pas limiter le temps; laisser le participant revenir en arrière ou aller de l'avant. Ne donner aucune autre instruction ou explication, sauf pour un exemple ou une répétition des instructions ci-dessus. L'expérimentateur ne doit pas indiquer le placement d'un terme quelconque. Il ne doit ni influencer ni aider le participant une fois que la séance a commencé. Lors de l'étude d'origine, peu de personnes ont eu du mal à comprendre ce qu'on attendait d'elles. Si cela se produit, il vaut souvent mieux prendre un autre participant.

3.2.2 Participants

Il faudrait demander des réponses à un nombre de participants aussi grand que possible (plus de 20 par groupe, par exemple) et de provenances aussi diverses que possible dans la région linguistique. Dans chaque groupe, il convient tout d'abord de comparer les résultats pour déterminer les différences perçues dans les régions linguistiques du pays.

3.2.3 Moyenne des résultats de l'échelle graphique

On peut donner à chaque réponse une valeur quantitative en mesurant la distance entre une des extrémités de la ligne et la marque inscrite par le participant. On peut ensuite calculer et porter sur un graphique les moyennes géométriques et arithmétiques et les écarts types.

3.3 Résultats actuels des évaluations des intervalles perçus entre descripteurs

Le Document [CCIR, 1986-90c] rend compte d'études effectuées en Allemagne selon la méthode que décrit le § 3.1. Les résultats sont consignés dans la Figure 1. Ces essais ont été faits avec environ 55 participants. On a aussi analysé les résultats en fonction des groupes d'âge (jeune/âgé) et de la région d'Allemagne d'origine (Nord/Sud). On n'a noté que des différences relativement insignifiantes.

Le Document [CCIR, 1986-90d] rend compte d'études effectuées en France selon la méthode que décrit le § 3.1. Les résultats sont donnés dans la Figure 1. Il y avait environ 50 participants, qui étaient quelque peu familiarisés avec les descripteurs des échelles de qualité et de dégradation à 5 notes du CCIR. Dans ce cas, la dispersion des résultats était plutôt plus faible que pour d'autres groupes linguistiques.

Les résultats obtenus aux Etats-Unis d'Amérique et en Italie ont déjà été exposés et décrits au § 3.1. On les trouve aussi sur la Figure 1. La tendance la plus évidente, qui ressort de trois des quatre graphiques des descripteurs qualitatifs (allemand excepté), se remarque à la partie inférieure. Les dénominations sont resserrées à de faibles intervalles. L'échelle à cinq points et quatre intervalles est en réalité une échelle à 4 points et 3 intervalles d'amplitudes inégales.

Les dénominations de l'échelle des dégradations s'échelonnent de façon assez régulière dans les trois langues, sans doute parce que la gêne est ressentie comme un phénomène essentiellement continu.

3.4 Evaluation subjective des rapports de protection pour des signaux de radiodiffusion en ondes décimétriques

Des études effectuées aux Etats-Unis d'Amérique ont porté sur un certain nombre de facteurs qui intéressent l'utilisation de la bande de télévision en ondes décimétriques par les services mobiles terrestres. Une de ces études [Jones, B.L., 1986] a essayé d'établir une corrélation entre le rapport signal utile/signal brouilleur et la qualité de l'image. Le document en référence décrit en détail la méthode d'essai. Pour des qualités d'image semblables ou différentes, les échelles graphiques et de rapports se sont toutes deux révélées appropriées.

Ces essais démontrent nettement que les téléspectateurs sont à présent plus exigeants pour la qualité de l'image que par le passé.

Le Tableau I donne un exemple de résultats des essais. Il montre les appréciations des participants au cours des deux essais distincts, pour un rapport de 28 dB sur l'image, les échelles de rapports et graphiques et le bon accord entre les deux. Pour être quotidiennement "acceptable", il faudrait que la qualité de l'image soit multipliée par trois.

TABLEAU I

U/B = 28 dB, décalage 10 kHz (525,24 MHz)

Sur l'échelle de rapports

où subjectivement "acceptable" = 100

	Moyenne géométrique	Ecart type géométrique
Experts	35,5	2,9
Non-experts	35,0	2,3
Tous les observateurs	35,3	2,4

Sur l'échelle graphique

Experts	= "Médiocre"
Non-experts	= "Pas tout à fait passable"
Tous les observateurs	= "Pas tout à fait passable".

3.5 Usage des descripteurs de l'échelle de qualité du CCIR

Les dénominations de la traditionnelle échelle de qualité de la Recommandation 500 du CCIR ont fait l'objet d'études approfondies. Dans plusieurs pays et dans plusieurs langues (France, Allemagne, Etats-Unis d'Amérique et Italie), elles ont été comparées pour préciser leur signification et l'étendue des intervalles qui les séparent. Les échelles graphiques qui en découlent ont aussi servi avec succès à évaluer subjectivement la qualité des images.

Dans le contexte de la TVHD et des méthodes à double stimulus, on a avancé que les dénominations traditionnelles ne s'appliquaient plus. Certains estiment qu'elles ne décrivent plus la qualité d'image perçue. Compte tenu, par exemple, du fait qu'il n'est plus question d'images de qualité inférieure ou dégradées, "mauvais", "médiocre" et même "assez bon" appartiennent-ils toujours à l'échelle? On n'a certainement pas à évaluer d'images mauvaises ou médiocres.

On a proposé et effectué une expérience où les dénominations étaient définies à la suite des appréciations de la qualité de l'image [CCIR, 1986-90e] (échelle des dénominations a posteriori). On demandait aux observateurs de placer sur leurs échelles quelques-unes ou toutes les dénominations qui décrivaient les images qui venaient de leur être présentées. Il s'agissait d'images NTSC 1125 lignes de qualité studio non dégradées.

Sans exception, les observateurs ont échelonné toutes les dénominations. Lorsqu'on leur parlait d'images "mauvaises" ou "médiocres", ils disaient qu'il n'y en avait pas; cependant, ils avaient placé ces vocables. Il est apparu que lorsqu'on comparait des séries de très bonnes images à des images de TVHD, ces dernières faisaient toujours descendre le long de l'échelle les réactions à de bonnes images et la signification des dénominations ne correspondait donc plus à la qualité de l'image.

On essayait, au moyen de cette étude, de se rendre compte si, lorsqu'on échelonnait les dénominations après avoir apprécié la qualité d'image, elles correspondaient mieux aux images présentées. Il n'en était rien. Les dénominations se situaient toujours de la même façon et ne correspondaient pas à la qualité de l'image. On peut en conclure qu'on ne posait pas la question qu'il fallait et que, lorsqu'on donne aux observateurs des dénominations à échelonner, ils le font en fonction de leur signification et sans tenir compte de leurs échelles de qualité d'image.

Il est proposé d'essayer une méthode binaire: demander d'abord à l'observateur s'il a vu des images qu'on pourrait caractériser par chacun des vocables en l'invitant à répondre par oui ou non. Puis échelonner les vocables pour lesquels la réponse a été oui. Peut-être obtiendra-t-on ainsi par contrainte une liaison plus véridique entre la signification des vocables et la qualité de l'image. On pourrait aussi demander à l'observateur de donner un vocable, ou de le choisir parmi un grand nombre de vocables, afin de décrire une qualité d'image connue et fixe (qu'il ne connaît pas).

4. Echelle catégorielle numérique

L'échelle catégorielle numérique se fonde sur la capacité des observateurs à formuler des appréciations selon des catégories dérivant d'une échelle linéaire. Comme les caractéristiques des catégories ne sont pas limitées par des adjectifs, on peut utiliser l'échelle pour des amplitudes très diverses.

Le nombre de points de l'échelle dépend des conditions d'expérience et de la gamme des attributs de la perception.

L'expérience a montré que, dans certains cas, 5 points suffisent, alors que parfois il faut des demi-points sur une échelle à 10 notes.

Il est avantageux de choisir une échelle à laquelle les observateurs sont habitués dans la vie courante. Dans certains pays d'Europe, par exemple, une notation sur 10 est usuelle à l'école.

Le travail est vite fait avec une échelle catégorielle numérique et il est facile à automatiser (10 boutons par exemple). Il existe des essais pour vérifier l'égalité des intervalles de l'échelle des appréciations [Edwards, 1957].

Si besoin est, on peut facilement convertir les nombres en échelons de catégories d'espacements égaux, au moyen d'un logiciel disponible qui met en jeu les modèles de Thurstone [Torgersen, 1958].

Avant de commencer vraiment une expérience, il est bon de procéder à quelques essais qui révèlent la gamme complète sans toutefois la spécifier. Cela permet aux observateurs de stabiliser dans la bonne gamme leur échelle individuelle. Pour chaque contexte, le nombre de ces essais dépend totalement de l'objectif de l'estimation. Mais pour avoir un contrôle statistique, il en faut au moins trois.

5. Echelonnement multidimensionnel

5.1. Introduction

Le Document [CCIR, 1986-90f] rend compte d'expériences au cours desquelles on demandait aux observateurs d'évaluer la ressemblance entre images plus ou moins affectées par le bruit, et les brouillages co-canal et canal adjacent. On a essayé de déduire des résultats un espace des perceptions à trois dimensions. Cela n'a été qu'un succès partiel, sans doute en raison d'effets aux limites. Les études se poursuivent.

5.2 Méthodes utilisant une métrique multidimensionnelle

Plusieurs chercheurs ont eu recours à des méthodes utilisant une métrique multidimensionnelle pour étudier des appréciations comparatives de stimuli en télévision [Linde et autres, 1981; Goodman et Pearson, 1979]. Un exercice classique de construction de métrique multidimensionnelle commence par de telles appréciations comparatives (catégorielle ou non, voir Recommandation 500) portant sur les ressemblances entre membres de couples de conditions. Ensuite, on se sert des appréciations de ressemblance pour définir des "distances" entre les conditions dans un espace de perceptions à n dimensions et on applique aux appréciations une des nombreuses méthodes classiques pour trouver et désigner les dimensions de cet espace [Shiffman et autres, 1981].

Ces méthodes peuvent contribuer à une étude à trois niveaux de la télévision. On emploie d'abord la métrique multidimensionnelle pour définir les dimensions de perception dans lesquelles varient les facteurs de conception et de transmission. Puis, dans l'espace de perception, les coordonnées des niveaux des facteurs peuvent servir à définir des relations entre paramètres objectifs et de perception. Et enfin, on peut établir une relation entre les dimensions de perception et les appréciations de la qualité ou la satisfaction du téléspectateur. Mais, pour le moment, les méthodes utilisant une métrique multidimensionnelle ont rarement servi à étudier la qualité de l'image de télévision. Il reste à étudier la valeur de ces méthodes et celle de l'approche générale. On trouvera une description complète de cette approche dans [Lupker et Hearty, 1987]. Son efficacité fait actuellement l'objet d'études au Canada.

5.3 Méthode à variables aléatoires multiples

Le projet ESPRIT 925 a mis en oeuvre une méthode analogue qui est à l'étude en Espagne [CCIR, 1986-90g]. Les expériences auront pour but de repérer, de grouper et d'interpréter les variables et les facteurs subjectifs pertinents qui affectent la qualité de l'image. Le matériel expérimental comprendra un questionnaire qui demandera d'exprimer des avis au sujet des séquences vidéo observées. Les questions seront par exemple les suivantes: Quel est l'attribut le plus prisé et le moins prisé? Les séquences sont-elles de qualité égale ou différente? En cas de dégradation de l'image, comment pourrait-on la corriger? Et quelles sont, par ordre d'importance, les caractéristiques qui donnent une haute qualité à l'image? (voir [CCIR 1986-90g] pour plus de détails). Il se peut que les variables qui ont le plus d'importance ou de poids soient les suivantes:

- Niveau sonore

- Image: contours (nets ou flous)
  brillance et éclairement
  précision
  mouvement (surtout horizontal)
  couleur
  bruit vidéo
  contraste
  papillotement
  contenu global et local
  agrément de la composition
  netteté des traits des visages
  expression des visages
  rapport d'éclairement entre visages et fond
  continuité de la séquence
  position des sujets

On peut assigner à ces variables des facteurs globaux comme:

- Contenu local et bruit

- Contenu et visage: fond

- Qualité globale (couleur, contraste, etc.)

- Expression des visages

- Netteté des visages

- Mouvement

et peut-être, en outre : contenu bruit et couleur.

On s'efforce actuellement d'améliorer et de compléter le questionnaire et de prouver que la méthode convient à diverses évaluations de la qualité de l'image de télévision.

5.4 Méthode d'essai orthogonale

Des études effectuées en Chine ont montré qu'en appliquant la méthode d'essai orthogonale aux évaluations subjectives de la qualité de l'image, on peut au moyen d'un petit nombre d'expériences et en tolérant certaines distorsions, arriver à une généralisation à partir d'un groupe de combinaisons de distorsions qui représentent les mêmes caractéristiques principales que pour un essai complet; on peut aussi vérifier, dans chaque combinaison, l'indépendance des distorsions. On obtient ainsi un groupe convenable de combinaisons de distorsions en vue d'évaluations subjectives de la qualité d'images affectées par cinq distorsions simultanées [CCIR, 1986-90h].

6. Echelle graphique avec triple présentation de stimuli

Les expériences ont montré que l'évaluation des images avec des échelles catégorielles ne fournissent qu'une échelle de classement. On peut aussi se tourner vers les méthodes utilisant une échelle graphique. Le Document [CCIR, 1986-90i] décrit une méthode utilisant une échelle graphique qui permet un contrôle du type d'échelle obtenue (échelle ordinale ou d'intervalles). Au cours d'une expérience avec triple présentation de stimuli, on demandait aux observateurs d'indiquer sous forme graphique où se situerait la qualité de l'image d'un écran de contrôle central, par rapport à celle observée sur deux autres écrans placés de part et d'autre. Les images étaient entachées de divers niveaux de bruit. On présentait toutes les combinaisons possibles d'une série de niveaux de bruit. Cette méthode a servi à montrer une échelle d'appréciation subjective risque de ne pas fournir une échelle d'intervalles.

7. Méthode par comparaisons deux à deux

7.1 Description générale

Cette méthode ne peut fournir qu'un classement des images ou des séquences en fonction de leur qualité subjective, qui se déduit d'évaluations faites pour toutes les paires possibles. Elle a l'avantage de faciliter l'examen de la transitivité des appréciations portées par plusieurs sujets et de la correspondance entre les critères qu'ils utilisent. Il n'est pas toujours évident que ces conditions soient remplies, notamment en cas de dégradations complexes de l'image. Si l'examen donne un résultat négatif, il faudra penser à une méthode d'appréciation multidimensionnelle (Echelonnement multidimensionnel ou analyse des facteurs).

La fiabilité de la méthode dépend du nombre d'images ou de séquences (au moins égal à 6) et du nombre de sujets.

7.2 Procédure de l'essai

Il s'agit de faire établir par N sujets un classement de n images ou séquences. On présente aux sujets toutes les paires possibles dans un ordre aléatoire. Pour chaque paire, ils indiquent quelle image ou séquence est la meilleure. Il y a au total

$$z = n\,(n-1)/2 \quad \text{paires}$$

Pour chaque sujet, les résultats des essais forment une matrice de préférence individuelle n x n. Un "un" dans la colonne t et sur la ligne s veut dire que l'image ou la séquence t est meilleure que celle de rang s, un "zéro" signifiant le contraire.

7.3 Examen de la transitivité individuelle (méthode du Zêta)

Il faut vérifier la transitivité des résultats fournis par tous les sujets. Un sujet a fourni une "triade non transitive" s'il a décidé qu'une image A était meilleure qu'une image B, B meilleure que C mais que C était meilleure que A. L'essai ne peut produire un classement que si les sujets ne portent que des appréciations transitives.

Dans une matrice de préférence, le nombre de triades non transitives est:

$$d = (n(n - 1)\ (2n - 1)/12 - 1/2) \sum_i D_i^2$$

où D_i est la somme de la ième colonne.

Au maximum d vaut:

$d_{max} = n\ (n^2 - 4)/24$ si n est pair

$= n\ (n^2 - 1)/24$ si n est impair

ζ est une mesure de la transitivité:

$\zeta = 1 - d/d_{max}$

Si ζ est égal à 1, il y a transitivité absolue. Si ζ est égal à 0, les appréciations sont totalement non transitives.

Les résultats transitifs pourraient aussi être aléatoires. En considérant la probabilité de la valeur calculée des triades non transitives (d) sous condition que la transitivité soit exclusivement aléatoire, on peut savoir si les appréciations portées par le sujet sont systématiquement transitives ou non. On se fixe d'abord une limite α de probabilité, assez faible, par exemple 0,05 (5%). Puis on calcule la valeur suivante:

$$x = \frac{8}{n - 4}\ (1/4 \binom{n}{d} - d + 1/2) + DF$$

$DF = n\ (n - 1)\ (n - 2)\ /\ (n - 4)^2$ (degrés de liberté)

Pourvu que $n > 6$, x (d) est une fonction à distribution en χ^2. χ^2 est une distribution de probabilité bien connue (Tableau II). On trouve dans ce tableau la valeur de χ^2 qui correspond à la valeur de α choisie et à la valeur de DF calculée. Si la valeur lue est inférieure à la valeur de x calculée, on admet qu'il y a transitivité systématique.

TABLEAU II

Distribution en χ^2

Degrés de liberté	$\alpha = 0,05$	$\alpha = 0,01$	$\alpha = 0,001$
1	3,84	6,63	10,83
2	5,99	9,21	13,82
3	7,81	11,34	16,27
4	9,49	13,28	18,47
5	11,07	15,09	20,52
6	12,59	16,81	22,46
7	14,07	18,48	24,32
8	15,51	20,09	26,13
9	16,92	21,67	27,88
10	18,31	23,21	29,59
11	19,68	24,73	31,26
12	21,03	26,22	32,91
13	22,36	27,69	34,53
14	23,68	29,14	36,12
15	25,00	30,58	37,70
16	26,30	32,00	39,25
17	27,59	33,41	40,79
18	28,87	34,81	42,31
19	30,14	36,19	43,82
20	31,41	37,57	45,32
21	32,67	38,93	46,80
22	32,92	40,29	48,27
23	35,17	41,64	49,73
24	36,42	42,98	51,18
25	37,65	44,31	52,62
26	38,89	45,64	54,05
27	40,11	46,96	55,48
28	41,34	48,28	56,89
29	42,56	49,59	58,30
30	43,77	50,89	59,70
40	55,8	63,7	73,4
50	67,5	76,2	86,7
60	79,1	88,4	99,6
70	90,5	100,4	112,3
80	101,9	112,3	124,8
90	113,1	124,1	137,2
100	124,3	135,8	149,4

7.4 Examen de la concordance entre les sujets

Il n'est raisonnable de calculer un classement commun que si les sujets fondent leurs appréciations sur les mêmes critères (c'est-à-dire si la concordance entre eux est systématique). Pour s'en assurer, on regroupe les résultats des essais dans une matrice dite agglomérée. A cette fin, on numérote toutes les paires d'images ou de séquences. Leur ordre est arbitraire. Ces nombres correspondent aux numéros des lignes de la matrice. Les numéros des colonnes correspondent aux numéros (arbitraires) des sujets. Dans cette matrice, l'élément X_{ij} est égal à 1 (ou à zéro) si le sujet j estime que la première image de la parie i est meilleure (ou pire) que la seconde.

La concordance entre les sujets peut être systématique ou aléatoire. On la suppose systématique si la probabilité de concordance effective est assez petite, sous réserve qu'elle ne soit qu'aléatoire. On se fixe d'abord une probabilité limite α qui soit assez faible, par exemple $\alpha = 0{,}05$ (5%). Puis on calcule la valeur suivante:

$$Q = \frac{\binom{n}{2}\left[\binom{n}{2} - 1\right] \sum_i (L_i - \bar{L})^2}{\binom{n}{2} \sum_j G_j - \sum_j G_j^2}$$

n: nombre d'images ou de séquences
N: nombre de sujets
L_i : somme de la ième ligne de la matrice agglomérée

$$\bar{L} = \Sigma L_i / \binom{n}{2}$$

ΣG_j : somme de la jème colonne de la matrice agglomérée.

On calcule aussi le nombre de degrés de liberté:

$$DF = \binom{n}{2} - 1$$

Pour la valeur de α choisie et la valeur calculée de DF, on lit la valeur correspondante de χ^2 dans le Tableau II. Si Q est supérieur à χ^2, on suppose que la concordance est systématique.

7.5 Calcul du classement

Sous réserve de la transitivité de tous les sujets et d'une concordance systématique entre eux, on peut tirer un classement des résultats des essais.

On calcule une matrice de préférence globale en additionnant les matrices individuelles. L'élément X_{ij} de cette matrice est égal à la fréquence de l'appréciation selon laquelle l'image j est meilleure que l'image i. Puis on fait la somme D_i des colonnes de la matrice. D_i est la fréquence de l'appréciation selon laquelle l'image i est la meilleure de toutes. L'ordre de ces sommes donne le classement des images ou des séquences.

8. Méthode d'évaluation de seuils de visibilité

Pour certaines mesures et notamment pour obtenir la meilleure précision lorsque l'on établit une correspondance entre des mesures objectives et l'évaluation de leur influence sur la qualité visuelle, il est intéressant de mesurer les seuils de visibilité de "facteurs de dégradation". Même si la méthode du double stimulus et échelle de dégradation peut apporter des informations à ce propos, il semble utile de préconiser l'emploi d'une méthode plus simple et plus performante appelée: "méthode du double stimulus en choix forcé". Le comportement de cette méthode, très souvent utilisée dans le cadre d'études psychophysiques, a été vérifié dans le contexte des études menées au CCIR; les résultats sont reportés dans [1986-90j].

La procédure utilisable aussi bien pour les matériaux naturels que synthétiques est basée sur la comparaison d'une séquence dégradée avec la référence correspondante. Dans chaque paire de séquences constituant une présentation, la position de la référence est aléatoire. La tâche des observateurs consiste seulement à indiquer laquelle des deux séquences de la présentation est dégradée. Le choix est dit forcé car les observateurs doivent toujours donner une réponse, même s'il y a doute.

Les niveaux de dégradation présentés doivent couvrir une gamme suffisamment large au-dessus et en dessous du seuil de visibilité estimé.

Le traitement des votes suit le protocole suivant: classiquement, la probabilité de bonnes réponses variant entre 50% (dégradation non vue, c'est-à-dire réponses données au hasard) et 100% (dégradation toujours vue), on fait une estimation du seuil de visibilité à 75% pour chaque observateur. Une fois le seuil de chaque observateur estimé, on calcule le seuil moyen et son intervalle de confiance associé. Un autre pourcentage peut être choisi pour la mesure d'un seuil moins rigoureux.

La stabilité de la méthode ainsi que son aptitude à fournir un véritable seuil de visibilité ont été vérifiées en comparaison avec la méthode à double stimulus utilisant une échelle de dégradation. Les résultats sont clairement en faveur de la méthode du double stimulus en choix forcé, quelle que soit la gamme de dégradations choisie mais il est préférable de présenter des niveaux de dégradation correctement répartis autour du seuil estimé.

9. Dispositions à prendre pour inclure une nouvelle méthode dans la Recommandation 500

D'après les éléments d'information actuels, s'il faut remplacer ou compléter les méthodes décrites actuelles dans la Recommandation 500, c'est la méthode utilisant une échelle de rapport (§ 2.1) qui a le plus de chances.

Une première expérience a montré que cette méthode est moins influencée par le contexte qu'une méthode d'échelle de qualité à un seul stimulus et une seconde expérience a laissé penser que la corrélation entre laboratoires serait bonne, pourvu que les références de qualité de l'image soient échelonnées ensemble.

Il faudrait que plusieurs laboratoires, travaillant dans des langues différentes, confirment ces perspectives.

Il importe aussi que la communauté des radiodiffuseurs donne ses directives quant à l'interprétation des résultats. Les radiodiffuseurs ont coutume d'utiliser des échelles de catégorie à 5 notes et on aurait besoin de connaître la relation entre les deux environnements.

Il convient également de poursuivre l'étude des méthodes utilisant une échelle numérique par catégories et utilisant une métrique d'échelonnement multidimensionnelle et graphique (§ 4 de la Recommandation 500), afin d'expliciter les avantages qu'elles présenteraient par rapport à d'autres méthodes possibles.

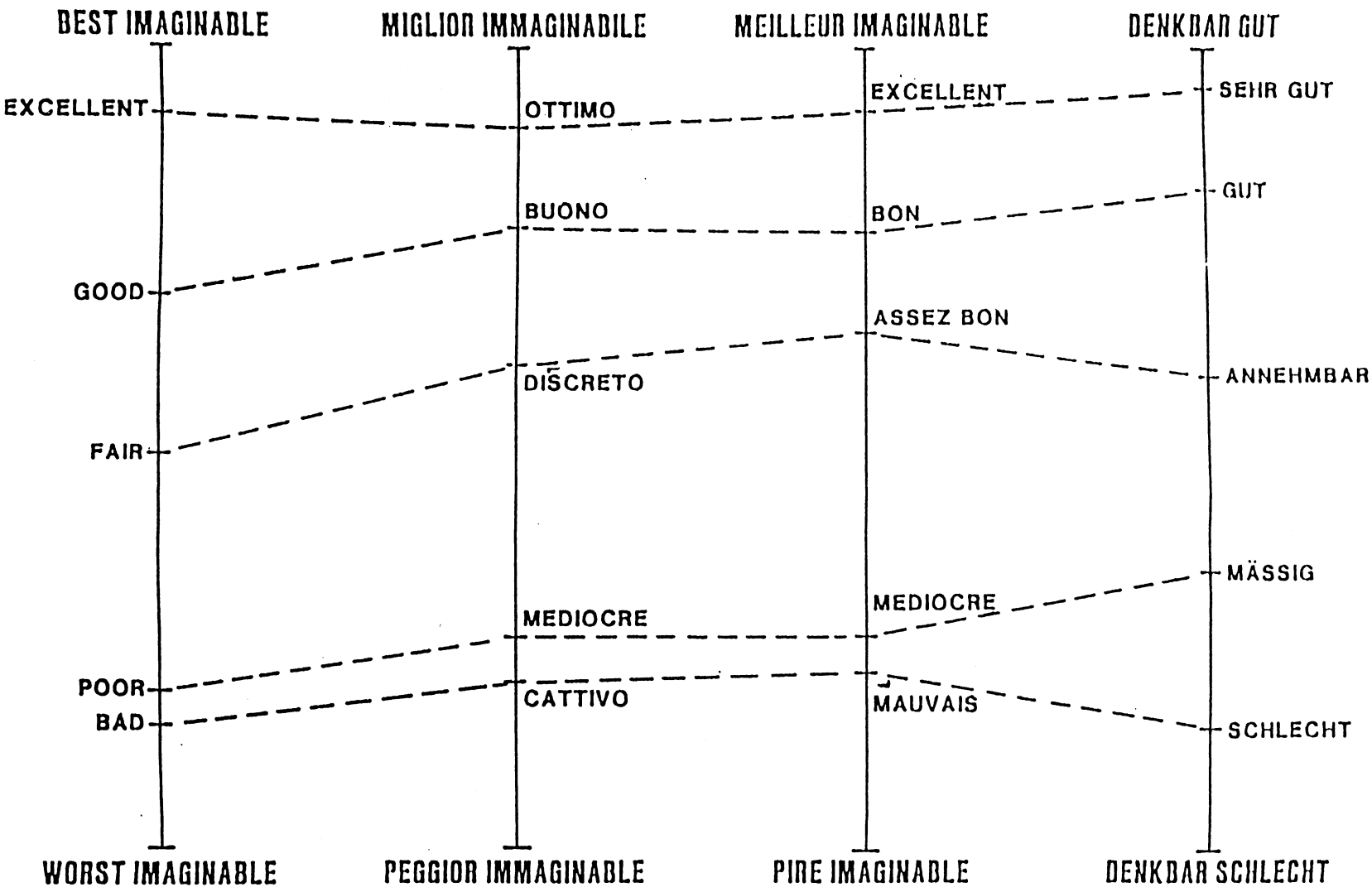

FIGURE 1a

Echelles graphiques des dénominations de qualité

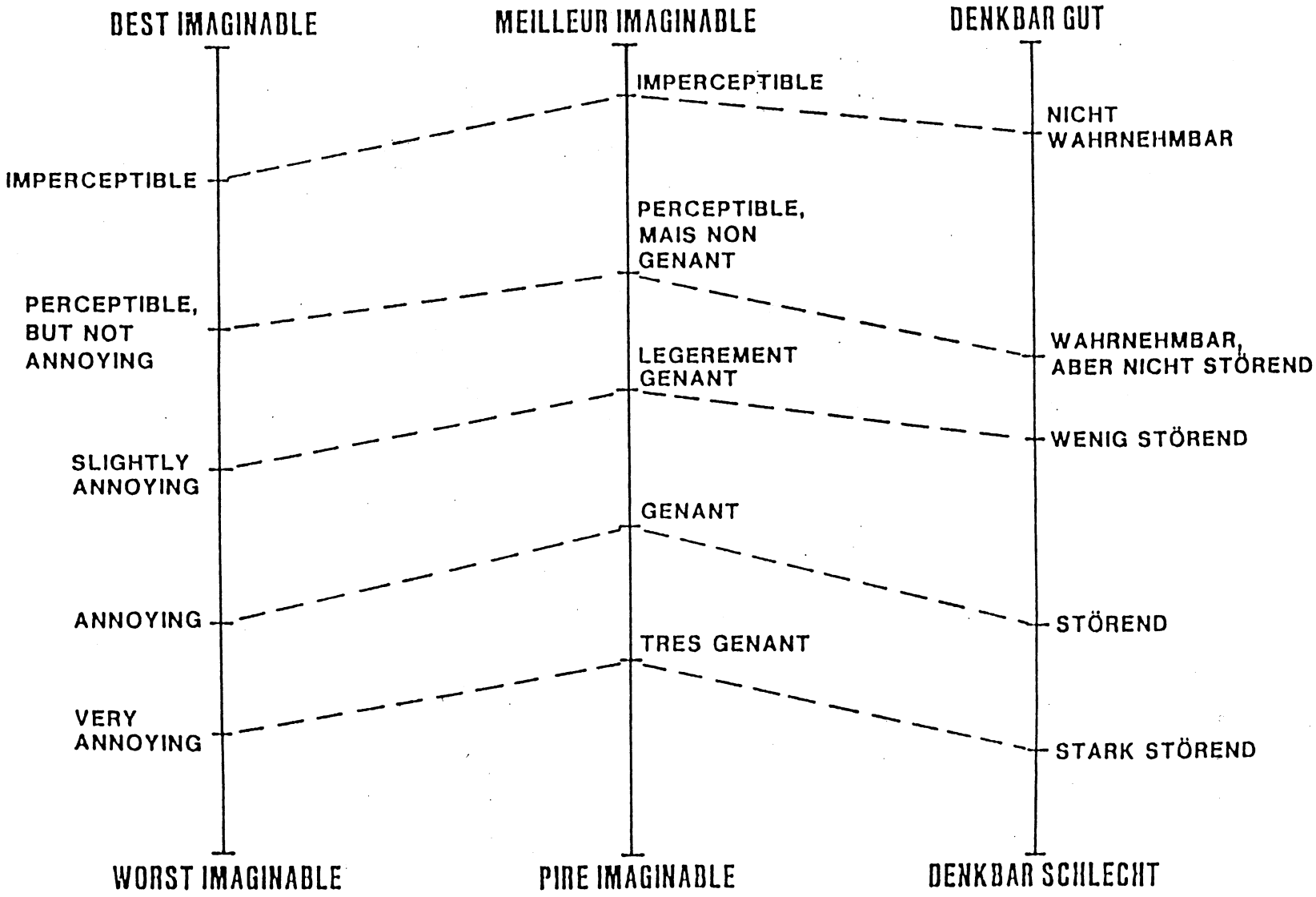

FIGURE 1b

Echelles graphiques des dénominations des dégradations

REFERENCES BIBLIOGRAPHIQUES

EDWARDS, A.L., [1957] - Techniques of attitude scale construction NX Appleton Century Crofts Inc.

GOODMAN, J.S. et PEARSON, D.E. [1979]. Echelonnement multidimensionnel des images de télévision à défauts multiples. IEEE Trans. Systems, Man. Cybernetics, 9, 353-356.

JONES, B.L. [11 juillet 1986]. Evaluation subjective des rapports de protection pour les signaux de radiodiffusion en bandes dm. Etude soumise à la FCC comme commentaires du NAB, General Docket, 85-172.

JONES, B.L. et McMANUS, P.R. [1986]. Echelonnement graphique des appréciations qualitatives. SMPTE J., Vol. 95, 11-86, 1166-1171.

LINDE, L., MARMOLIN, H. et NYBERG, S. [1981]. Résultats visuels de l'échantillonnage dans le traitement numérique de l'image - expérience pilote. IEEE Trans. Systems, Man. Cybernetics, 11, 201-207.

LUPKER, S. et HEARTY, P. [1987]. Evaluating the effects of multiple sources of impairment in TV signals.
Proc. 3rd International Colloquium on Advanced Television Systems: HDTV 87, Ottawa, Canada.

SCHIFFMAN, S.S., REYNOLDS, M.L. et YOUNG, F.W. [1981]. Introduction à l'échelonnement multidimensionnel. New York, Academic Press.

TORGERSEN, W.S., [1958] - Theory and methods of scaling NY John Wiley & Sons.

Documents du CCIR

[1986-90]: a. 11/379 (USA); b. GTI 11/4-123 (France); c. GTI 11/4-137 (Allemagne (République fédérale d')); d. GTI 11/4-147 (France); e. GTI 11/4-160 (Rapporteur principal, GTI 11/4), f. GTI 11/4-145 (Canada); g. 11/158 (Espagne); h. 11/144 (République populaire de Chine); i. GTI 11/4-141 (Allemagne (République fédérale d')); j. 11/463 (France);

ANNEXE I

Caractéristiques des méthodes d'évaluation pendant l'émission des programmes

Références	OIRT [CCIR, 1966-69 a]		Canada [CCIR, 1966-69 b]
Observateurs			
Catégorie	Spécialiste		Spécialiste
Nombre	1 ou 2		1 ou 2
Echelle d'évaluation			
Type	Dégradation	Qualité	Dégradation
Nombre de notes	6 (Note 1)	6 (Note 2)	5 (Note 3)
Image			
Type	Programmes de télévision		Programmes de télévision
Conditions d'observation			
Rapport de la distance d'observation à la hauteur de l'image	4 à 6		4 à 6
Angle de vision avec une ligne perpendiculaire au récepteur de contrôle			$\leq 30°$
Luminance sur l'écran pour le blanc de référence (cd/m^2)			70 ± 7
Chromaticité de l'écran pour le blanc de référence			Illuminant *D*
Luminance de l'écran du tube éteint	Adapté à l'éclairage ambiant		Aussi faible que possible
Luminance de «l'encadrement lumineux» (cd/m^2)			$10{,}5 \pm 3{,}5$ (Note 14)
Chromaticité de «l'encadrement lumineux»			Illuminant *D*

Note 1. – ***Echelle de dégradation*** de 6 degrés

1 imperceptible
2 juste perceptible
3 nettement perceptible, mais non gênant
4 légèrement gênant
5 nettement gênant
6 inutilisable

Note 2. – ***Echelle de qualité*** de 6 degrés

1 excellent
2 bon
3 assez bon
4 médiocre
5 mauvais
6 très mauvais

Note 3. – ***Echelle de dégradation*** de 5 degrés

1 imperceptible (implicite)
2 perceptible
3 apparent
4 gênant
5 inutilisable

REFERENCES BIBLIOGRAPHIQUES

Documents du CCIR

[1966-1969]: a. XI/46 (OIRT); b. XI/146 (Canada).

ANNEXE II

Renseignements supplémentaires sur les procédures d'essai

La présente annexe décrit et examine quelques procédures relatives à l'organisation d'essais d'évaluation subjective. La Recommandation 500 donne beaucoup d'informations sur l'évaluation subjective et les méthodes d'essai les plus couramment utilisées. Le § 1 de cette annexe décrit des procédures additionnelles et nouvelles; le § 2 décrit la technique de l'échantillon virtuel; le § 3 donne quelques informations sur les résultats obtenus avec certaines procédures dans des situations comparables; le § 4 donne quelques éléments d'information supplémentaires et le § 5 expose les différents arguments concernant le choix d'une méthode appropriée pour les diverses situations.

1. Description des procédures

Ce paragraphe présente plusieurs procédures qui s'ajoutent à celles décrites dans la Recommandation 500, y compris de nouvelles procédures destinées à résoudre de nouveaux problèmes. Les descriptions concernent les échelles de notation, la détermination de la séquence de présentation, la définition des conditions avec dégradation, etc. La validité de ces données n'est pas nécessairement limitée à l'une des procédures particulières décrites ci-dessous.

1.1 *Procédure utilisant l'échelle de qualité avec ancrage direct**

Pour étudier le rôle de l'image de référence, une nouvelle méthode a été définie [Kretz et Sallio, 1981]. Dans cette méthode, l'échelle de qualité à 5 notes de la Recommandation 500 est utilisée et la procédure est la même que celle de la méthode de l'UER (voir plus bas le § 1.2): une image de référence est présentée avant chaque image à juger et on indique aux observateurs que les images sont à juger par rapport à l'image de référence, et que celle-ci correspond à la note 5 («excellent»). Avec cette procédure, on obtient un ancrage direct des notes au sommet de l'échelle en utilisant l'échelle d'évaluation de la qualité.

1.2 *Procédure utilisant des séquences d'images animées*

La Recommandation 500 précise les conditions d'évaluation subjective de la qualité des images fixes et des images animées. Toutefois, peu d'études dont on ait connaissance traitent des principales caractéristiques de mesure de la qualité subjective des séquences d'images animées. Ces caractéristiques comprennent:

- les éléments permettant à l'observateur de prendre sa décision sont souvent fugitifs;
- l'observateur peut difficilement analyser et évaluer tous les éléments de la séquence en une seule observation;
- la perception de certaines dégradations peut être différente sur images fixes et sur images animées.

Dans une première tentative de définir une procédure appropriée, une étude [CCIR, 1982-86a], utilisant l'échelle de dégradation avec une image de référence (méthode UER) a examiné plusieurs modes de présentation des séquences d'images animées. La dégradation utilisée était du type enregistrements successifs sur magnétoscopes.

Les conclusions préliminaires ci-après peut être tirées sur la base des notes moyennes et des écarts types. Premièrement, les séquences d'une longueur de plus de 10 s semblent être trop longues. Deuxièmement, la répétition des séquences, dans le but de permettre à l'observateur de mieux analyser et évaluer les scènes, ne semble pas améliorer la qualité des évaluations. La conclusion de cette étude préliminaire semble être qu'une présentation unique d'environ 8 à 10 s est préférée. En outre, les résultats montrent que les évaluations des mêmes dégradations sur des séquences d'images fixes et sur des séquences d'images animées peuvent différer de manière significative.

Les conclusions qui précèdent sont préliminaires et sont fondées sur les résultats obtenus avec un seul type de dégradation; les résultats obtenus avec des dégradations dépendantes du mouvement pourraient différer de manière significative. Les études sur cette question importante doivent être poursuivies d'urgence.

---

* Une forme d'ancrage est toujours implicite dans toutes les procédures en ce sens qu'il est nécessaire de normaliser le processus de notation. Ici, le terme «ancrage direct» correspond à un ancrage *explicite*. L'ancrage «indirect» correspond à la normalisation des phénomènes d'adaptation [Corbett, 1970], au moyen de la dynamique des dégradations d'une séance d'essai.

2. Méthode de l'échantillon virtuel

2.1 Les principales sources d'erreurs dans les essais subjectifs sont essentiellement au nombre de deux:

- les erreurs aléatoires (stochastiques), et
- les erreurs systématiques.

Pour autant que les conditions objectives de l'essai aient été normalisées, la nature des erreurs dépend seulement des paramètres utilisés pour la conception des essais (nombre d'observateurs, images, méthode utilisée, etc.).

Les erreurs stochastiques sont très faciles à identifier. Lorsqu'il s'agit de déterminer une relation entre la valeur d'une distorsion et la qualité d'image, ces erreurs entraînent une certaine dispersion des notes moyennes expérimentales autour de la courbe lissée (obtenue, par exemple, par application de la méthode des moindres carrés). La normalisation des méthodes d'essai a généralement pour effet de réduire les erreurs stochastiques mais, du fait de la nature de ces erreurs, il est possible de les réduire encore par un procédé statistique d'établissement de moyenne (le lissage au moyen d'une fonction mathématique constitue un de ces procédés d'établissement de moyenne).

Les erreurs systématiques sont difficiles à reconnaître car elles n'ont pratiquement aucune corrélation avec les erreurs aléatoires. Dans la plupart des cas, elles ont pour effet de déplacer la courbe mentionnée plus haut (biais) et/ou de modifier sa pente. Une fois qu'on a introduit un certain degré d'erreurs systématiques dans les résultats expérimentaux, il n'est plus possible de les éliminer par des méthodes statistiques.

2.2 L'«échantillon virtuel» est constitué par un nombre relativement grand d'observateurs (par exemple, 50 ou plus) et par un grand nombre d'images (par exemple, 40 ou plus). On l'appelle «échantillon virtuel» parce que cet échantillon n'est pas utilisé intégralement lors des évaluations; on s'en sert comme d'une population dans laquelle on prélève successivement des échantillons plus petits et plus faciles à utiliser.

Prenons comme exemple le cas où il s'agit d'établir une relation entre des valeurs de distorsion et la qualité d'image sans oublier que le but de la technique de l'échantillon virtuel est d'utiliser différents échantillons d'observateurs et d'images pour différentes conditions d'essai. La méthode de l'échantillon virtuel serait alors appliquée de la manière suivante:

- on choisit un certain nombre de conditions d'essai (par exemple, 8 à 10 valeurs de distorsion);
- on forme un certain nombre de groupes, composés chacun d'un maximum de 2 ou 3 conditions d'essai non contiguës;
- pour chaque condition expérimentale, on choisit dans l'ensemble total (échantillon virtuel) un échantillon aléatoire de 5 ou 6 images d'essai; chaque groupe de condition d'essai comprendra par conséquent 2 ou 3 de ces ensembles d'images d'essai. Pour chaque groupe, on choisit un échantillon de 8 à 10 observateurs, prélevé sur l'ensemble total (échantillon virtuel). On obtient ainsi des images d'essai différentes pour des conditions d'essai différentes dans le même groupe;
- pour chaque groupe de conditions d'essai, on organise une séance d'essai en suivant la procédure décrite au § 8.4 de l'Appendice à la Recommandation 500-3, Recommandations et Rapports du CCIR, Vol. XI-1, Dubrovnik, 1986;
- on calcule les notes moyennes et on les ajuste par la méthode des moindres carrés, en appliquant une fonction convenable (par exemple, la «fonction logistique»);
- on effectue un essai statistique pour vérifier la qualité de l'ajustement; le résultat final de l'expérience est la courbe lissée ainsi obtenue.

Si l'on désire obtenir la courbe lissée avec une meilleure précision, on peut répéter une nouvelle expérience; les notes moyennes correspondantes doivent être moyennées avec les notes précédentes avant d'effectuer un nouvel ajustement par les moindres carrés.

Dans l'ensemble, il y a accord sur le fait que cette technique est applicable pour les dégradations complexes, du moins en ce qui concerne les observateurs. Cependant, certaines administrations estiment que cette technique ne serait peut-être pas économique dans le cas d'une dégradation unique.

3. Résultats d'expériences directement comparables

Ce paragraphe décrit des résultats obtenus au moyen de différentes méthodes d'évaluation sur le même matériel expérimental (images et dégradations). Naturellement, l'utilisation d'une méthode dans une situation particulière doit tenir compte de nombreux paramètres et les conclusions de telles expériences, décrites dans ce paragraphe, ne prennent en compte que quelques-uns de ces facteurs. Ce point est examiné en détail au § 5.

3.1 *Comparaison des résultats obtenus par la méthode à un stimulus et la méthode de l'UER*

Les deux procédures décrites aux § 8.1 et 8.2 de l'Appendice à la Recommandation 500-3, Recommandations et Rapports du CCIR, Vol. XI-1, Dubrovnik, 1986 ont été soumises à une série d'expériences comparatives [CCIR, 1978-82a, b, et c; Sallio et Kretz, 1982]. Différents types de dégradations ont été testés: filtrage analogique, bruit additif (dans deux contextes différents), bruit de transmodulation, flottement de contour (dans deux contextes différents), cumul de bruit additif et de flottement de contour, erreurs de transmission à 34 Mbit/s (pour deux dynamiques de dégradations). Vingt groupes indépendants de dix observateurs non experts (un groupe pour chaque type de dégradation et chaque procédure) ont participé aux essais (au total, 46 séances ont été nécessaires pour chaque méthode). On a comparé les résultats obtenus en analysant les valeurs des notes moyennes et des écarts types séparément pour chaque distance d'observation. On en a dégagé les conclusions suivantes:

- Les deux méthodes se traduisent par des courbes reflétant la relation objective-subjective qui sont de forme très similaire. On constate une translation entre les courbes obtenues selon les deux méthodes, la courbe d'évaluation de la qualité étant assez en dessous de la courbe en dégradation (courbe pour la méthode de l'UER). A mi-échelle, (note 3) les écarts types des notes atteignent un maximum et sont très voisins pour les deux procédures (à $6H$, on obtient 0,84 pour la méthode à un stimulus et 0,79 pour la méthode de l'UER).
- La procédure utilisant l'échelle de dégradation et une image de référence aboutit à une note moyenne pour les images de référence très voisine de la meilleure note (4,88 en moyenne), montrant un bon ancrage des jugements, au sommet de l'échelle.
- La procédure utilisant l'échelle de qualité conduit à une note moyenne pour les images de référence qui diffère de la note maximale d'une valeur voisine de 0 à une valeur voisine de 1 (4,56 en moyenne). Cela semble être dû au caractère absolu des appréciations.
- Les deux méthodes testées sont sensibles au contexte et à la dynamique des dégradations présentées lors de la même séance. Ces deux phénomènes subjectifs ont des effets identiques pour les deux méthodes. Il semble donc important, lors de la présentation des résultats, de décrire précisément les conditions expérimentales (dynamique des dégradations présentées dans une même séance, contexte de la séance), ce qui permettrait une meilleure compréhension des résultats.
- Dans la gamme de dégradations comprise entre «imperceptible» et «perceptible mais non gênant» (autour de la note 4,5), la procédure qui utilise l'échelle de dégradation et une image de référence conduit à des écarts types de notes 1,4 fois inférieurs à ceux que donne la procédure qui utilise l'échelle de qualité; la première procédure semble donc donner une meilleure précision et pourrait permettre de diviser par deux le nombre des notes vers le sommet de l'échelle.

Ces résultats suggèrent qu'il est possible d'obtenir une transformation des notes moyennes de qualité en notes moyennes de dégradation obtenues selon la méthode de l'UER en décalant les valeurs expérimentales d'une quantité égale à la valeur associée à la dégradation résiduelle (note moyenne associée aux images non dégradées). On suggère que la transformation de notes moyennes de dégradation en notes moyennes de qualité s'effectue en décalant les valeurs expérimentales d'une demi-note, bien que cette valeur ne soit pas complètement stable. Ces transformations ne peuvent être appliquées au bas de l'échelle, compte tenu du fait que l'échelle est bornée.

3.2 *Comparaisons des résultats obtenus avec d'autres méthodes*

Pour étudier plus en détail le rôle des images de référence, de l'ancrage et de l'échelle de notation, on a testé plusieurs méthodes sur le même type de dégradation [CCIR, 1978-82b]. Cette étude a consisté à comparer les résultats fournis par les deux méthodes citées plus haut (§ 3.1) et les résultats fournis par plusieurs autres méthodes. Les aspects suivants ont été examinés:

- utilisation d'une échelle de dégradation à 5 notes et d'une échelle de dégradation continue, de l'échelle de qualité à 5 notes et d'une échelle de qualité continue, toutes avec une image de référence pour un ancrage direct (il s'agit en fait de la procédure de l'UER avec des échelles différentes);
- utilisation d'une échelle continue de qualité, avec application d'une procédure à double stimulus, voisine de celle décrite au § 8.3 de l'Appendice à la Recommandation 500-3, Recommandations et Rapports du CCIR, Vol. XI-1, Dubrovnik, 1986;

L'étude a permis de tirer les conclusions suivantes:

- il est possible d'obtenir un bon ancrage dans la gamme des bonnes qualités, avec une présentation du type UER, mais en utilisant une échelle de qualité à 5 notes et en avertissant les observateurs que l'image de référence doit correspondre à la note «excellent»;

- en comparant les résultats fournis par des méthodes qui diffèrent seulement par l'emploi d'une échelle de notation discrète ou d'une échelle de notation continue, on constate que ni l'échelle continue de qualité, ni l'échelle continue de dégradation ne donnent plus d'informations que les échelles à 5 notes recommandées (moyennes et écarts types comparables);
- la méthode à double stimulus ne permet pas d'avoir un ancrage direct, la note moyenne pour la référence n'est pas proche du sommet de la gamme supérieure des notes. Les écarts types obtenus avec cette méthode ne sont pas significativement inférieurs à ceux que l'on mesure avec la méthode de l'UER;
- l'utilisation d'échelles continues soulève des problèmes: certains observateurs (non techniciens) ont des difficultés à appliquer ces échelles et l'analyse et la présentation des résultats deviennent plus compliquées.

3.3 *Comparaisons de résultats obtenus avec des échelles de qualité à 5 et 6 notes et avec une échelle de dégradation à 6 notes*

La comparaison des résultats obtenus avec l'échelle de qualité à 5 notes et avec l'échelle de dégradation à 6 notes a été décrite dans [Allnatt et Corbett, 1974] et réexaminée récemment surtout du point de vue de la qualité près du seuil de visibilité [Allnatt, 1980]. La procédure était la même sauf en ce qui concerne les échelles utilisées. L'échelle de dégradation à 6 notes est différente de celle de l'actuelle Recommandation 500. Deux types de dégradation ont été considérés, l'un est un écho sans distorsion de 2 µs en télévision monochrome à 625 lignes, le second est un flou dégradant des photographies. Les résultats ont été analysés uniquement en terme de note moyenne d'opinion. La principale conclusion de cette étude est que l'échelle de dégradation ne présente pas une meilleure sensibilité aux dégradations en dessous du seuil de visibilité. Toutefois, il convient de noter que cette expérience ne correspond pas à une comparaison de résultats que l'on aurait obtenus en comparant la méthode à un stimulus et la méthode de l'UER.

4. Quelques faits expérimentaux supplémentaires

On dispose de faits expérimentaux provenant d'autres sources, dont certains diffèrent de ceux indiqués au § 3. Il existe un volume considérable de résultats, dans le cas de la méthode à un stimulus (méthode de notation de la qualité), relatifs à toutes les propriétés importantes de cette méthode.

Les résultats confirment les possibilités d'application de la transformation imp (voir par exemple [Macdiarmid et Allnatt, 1978]; voir aussi l'Annexe II) sous la forme d'une loi d'addition des dégradations subjectives. Cette loi est utilisée pour compenser l'effet de la dégradation résiduelle dans les analyses et elle n'agit pas de la même façon que la transformation faisant appel à la translation des notes moyennes, préconisées au § 3.1.

Pour ce qui est de la méthode de notation de la qualité à double stimulus, on a constaté que les écarts types des différences entre les paires de notes d'une même présentation sont inférieurs à ceux des notes individuelles, lorsque les dégradations sont peu importantes. Une transformation est nécessaire pour déterminer l'écart type équivalent, et dans une échelle de notation à 5 notes, pour une dégradation nulle, des valeurs d'environ 0,13 ont été constatées avec une gamme relativement étendue de dégradations utilisant un bruit aléatoire [White et Allnatt, 1980], et 0,35 pour l'évaluation d'un codec numérique de haute qualité [CCIR, 1978-82d]. Au cours d'autres expériences réalisées en télévision numérique [IBA, 1981], on a obtenu la valeur 0,22; des valeurs comparables ont été obtenues au cours d'expériences analogues [Kretz et Sallio, 1981] (0,25 à 4*H* et 0,45 à 6*H*, pour une dégradation nulle).

La comparaison des résultats des mesures pour un codec numérique [CCIR, 1978-82d], par la méthode à double stimulus, et des résultats obtenus avec la méthode à un stimulus avec les dégradations à évaluer «plongées» dans d'autres dégradations variant dans une large dynamique, confirme la conclusion de White et Allnatt, [1980] selon laquelle les effets d'adaptation dus à l'ancrage indirect sont considérablement réduits lorsqu'on utilise la méthode à double stimulus, comparativement à la méthode à un stimulus.

Des résultats obtenus au moyen d'une échelle de dégradation montrent qu'il n'est pas souhaitable, lorsque l'on désire mesurer la visibilité de dégradations distinctes, de les présenter dans une même séquence [CCIR, 1982-86b]. Cette manière d'opérer peut introduire un biais, les observateurs ayant tendance à comparer l'effet des différentes dégradations. Il paraît donc préférable dans le cas de mesure de visibilité de dégradations distinctes, de ne faire juger qu'un seul type de dégradation par séquence de présentation.

5. Discussion

La spécification détaillée d'une procédure vise à réduire au minimum les variations aléatoires des résultats qui ne tiennent pas à des différences systématiques constatées entre les diverses populations d'observateurs lorsque, par exemple, les résultats d'essais indépendants sont comparés ou combinés. Toutes les procédures décrites visent à cette même fin, mais il existe d'autres facteurs qui, pense-t-on, donnent du poids à certaines procédures. Ces facteurs ont trait au degré de discrimination des résultats qu'il est utile de retenir, et au degré de complexité et de perfectionnement de la procédure d'essai. Plus cette procédure est complexe et perfectionnée, plus il est probable qu'elle demandera du temps et qu'elle sera coûteuse. La précision désirée et le gain réalisé du point de vue des résultats obtenus sont au centre de la discussion sur le choix de la procédure.

Les procédures décrites au § 8 de l'Appendice à la Recommandation 500-3, Recommandations et Rapports du CCIR, Vol. XI-1, Dubrovnik, 1986; retiennent une combinaison de l'échelle de qualité ou de l'échelle de dégradation, avec utilisation normale d'une image de référence, indiquée ou non en tant que telle ou seulement l'ancrage indirect par la gamme des dégradations. La procédure 8.1 est la méthode la plus simple du point de vue de l'organisation; l'analyse des résultats pour les § 8.1, 8.2 et le § 1.1 de la présente Annexe présente environ la même complexité; l'organisation des § 8.2 et 8.3 de l'Appendice et du § 1.1 de la présente Annexe est à peu de chose près la même, mais l'analyse des résultats pour le § 8.3 demande davantage de temps. La méthode 8.4 exige généralement un nombre plus important de séances que les autres procédures et elle est conçue pour réduire les erreurs systématiques. Une comparaison des résultats obtenus par certaines des méthodes pour certaines dégradations est donnée au § 3 ainsi que d'autres faits expérimentaux au § 4 de la présente Annexe.

Il va de soi que chaque méthode présente certains avantages et qu'il n'est pas facile de faire un choix. Il est impossible de rendre entièrement justice aux arguments présentés dans un Rapport tel que celui-ci, mais les principaux arguments qui ont influencé les chercheurs travaillant dans ce domaine sont essentiellement les suivants.

Le choix de la procédure est lié au choix de l'échelle de notation, selon qu'elle doit être continue ou discrète, et à la façon dont les observateurs doivent être priés de faire une utilisation correcte de l'échelle.

Pour ce qui est de la valeur relative de l'échelle de qualité et de l'échelle de dégradation, certains estiment que la notion de «qualité» est davantage liée à l'intérêt des observateurs et, en outre, qu'elle est avantageuse car si une «dégradation» améliore effectivement l'image, cela se trouve reflété dans les résultats. Par contre, d'autres estiment que l'échelle de dégradation est plus facile à interpréter et qu'elle présente l'avantage de permettre de mesurer un seuil de perception (entre la note 4 et la note 5 sur l'échelle de dégradation). Il semblerait que l'opinion du public se partage également quant à la faveur rencontrée vis-à-vis des deux échelles. On a montré qu'il peut exister un moyen, dans certains cas ou dans tous les cas, d'associer les deux axes sémantiques par une formule appropriée, et les travaux se poursuivent dans cette voie.

Les arguments en faveur d'une échelle continue sont que le temps supplémentaire consacré à l'organisation et à l'analyse est parfois justifié étant donné qu'une discrimination fine est à la fois possible et nécessaire. Les arguments en faveur d'une échelle discrète sont qu'il n'est pas possible d'obtenir de meilleurs résultats avec une échelle continue, notamment parce que les non-spécialistes ne font pas dans la pratique plus de distinctions que l'échelle à échelons ne le permet.

La nécessité de certaines formes d'ancrage est reconnue par tous mais on estime qu'elle peut être réalisée de différentes façons. Certains chercheurs suggèrent que, pour des dégradations qui présentent un intérêt pour une large gamme de valeurs, il n'est pas nécessaire de répéter une image de référence spécifique parce qu'une large gamme normalisée de dégradations incite les observateurs à s'orienter eux-mêmes correctement sur l'échelle. Pour des expériences avec de très faibles dégradations, une image d'ancrage devrait être utilisée sans être indiquée en tant que telle, pour ne pas rendre les conditions d'observation trop artificielles. Il se peut que la procédure à double stimulus présente un avantage dans la mesure de petites dégradations, comme celles des systèmes futurs. D'autres chercheurs prétendent que l'utilisation régulière et spécifiée d'une image de référence (qualité élevée) aide les observateurs à s'orienter eux-mêmes sur l'échelle et que les résultats des expériences le démontrent, notamment pour des dégradations peu importantes. Une autre méthode que l'on estime valable consiste à utiliser deux images de référence sans l'indiquer (qualité haute et qualité basse).

En ce qui concerne la méthode de l'UER, on a trouvé dans un cas une note aussi basse que 4,63. De son côté, l'Australian Broadcasting Commission [1981] a obtenu une note pour la référence de 4,42, dans un cas particulier. La SMPTE a fait des évaluations qui utilisaient des références NTSC et RGB, avec des observateurs qui étaient des employés de l'industrie de la radiodiffusion; elle a obtenu pour l'image de référence une note de 4,7. Certains chercheurs pensent que des valeurs faibles peuvent être dues, entre autres, à un contrôle insuffisant de la méthode ou à une qualité instable de l'image de référence. Dans des cas de ce genre, il semble nécessaire de corriger les dégradations résiduelles, comme dans les méthodes à un stimulus. Il pourrait être utile d'entreprendre de nouvelles études à propos d'autres méthodes.

Dans le cadre d'une évolution vers la rationalisation des méthodes, il serait possible au cours de la prochaine période d'études du CCIR de rassembler les principaux éléments des procédures et de limiter les possibilités seulement à certaines parties. Il est encourageant de noter que récemment dans une situation pratique (l'étude de la relation entre les dégradations et les différentes fréquences d'échantillonnage numérique [CCIR, 1978-82e, f, g]), situation dans laquelle on avait appliqué des règles de procédures précises, on est parvenu pratiquement aux mêmes notes moyennes en réalisant des essais entièrement indépendants à l'aide de procédures différentes (voir les § 8.2 et 8.3 de l'Appendice à la Recommandation 500).

REFERENCES BIBLIOGRAPHIQUES

ALLNATT, J. W. [1980] Subjective assessment method for television digital codecs, *Electron. Lett.*, **16**, 450-451.

ALLNATT, J. W. et CORBETT, J. M. [août 1974] Comparisons of category scales employed for opinion rating, *Proc. IEE*, Vol. 121, **8**, 785-793.

AUSTRALIAN BROADCASTING COMMISSION [1981] Tests of subjective impairment due to random noise (Rapport à paraître).

CORBETT, J. M. [mars 1970] Effect of observer adaptation on the results of television quality-grading tests, *Proc. IEE*, Vol. 117, **3**, 512-514.

IBA [1981] Subjective assessment of television quality. Experimental and development Report 114/81.

KRETZ, F. et SALLIO, P. [septembre-octobre 1981] Comparaison de deux méthodes d'évaluation de la qualité subjective des images: rôle des images de référence, de l'ancrage et de l'échelle de notation. *Radiodif.-Télév.*, Vol. 4/5, **69**, 37-42.

MACDIARMID, I. F. et ALLNATT, J. W. [juin 1978] Performance requirements for the transmission of the PAL coded signal. *Proc. IEE*, Vol. 125, **6**, 571-580.

SALLIO, P. et KRETZ, F. [avril 1982] Comparaison de deux méthodes d'évaluation de la qualité subjective des images de télévision. Représentation des résultats par une unité unique. *Rev. de l'UER* (Technique), **192**, 59-69.

WHITE, T. A. et ALLNATT, J. W. [1980] Double-stimulus quality rating method for television digital codecs, *Electron. Lett.*, **16**, 714-716.

Documents du CCIR

[1978-82]: **a.** 11/257 (France); **b.** 11/258 (France); **c.** 11/71 (France); **d.** 11/288 (Royaume-Uni); **e.** 11/285 (Royaume-Uni); **f.** 11/292 (Etats-Unis d'Amérique); **g.** 11/343 (Japon).

[1978-86]: **a.** 11/306 (France); **b.** 11/111 (France).

BIBLIOGRAPHIE

BENNETT, D. [octobre 1981] SMPTE component-coded digital video picture quality assessment. *SMPTEJ*, Vol. 90, **10**, 960-967.

BERNATH, K., KRETZ, F. et WOOD, D. [avril 1981] Méthode de l'UER pour l'organisation des essais subjectifs d'évaluation de la qualité des images de télévision. *Rev. de l'UER* (Technique), **186**, 66-75.

FISHER, R. A. et YATES, F. [1970] *Statistical Tables for Biological, Agricultural and Medical Research*, Oliver and Boyd, Edimbourg, Ecosse, Royaume-Uni.

ISHIHARA, S. [1949] *Tests for Colour Blindness.* H. K. Lewis, Londres, Royaume-Uni.

KRETZ, F. [septembre-octobre 1981] Représentation unifiée des résultats d'essais subjectifs (correction pour dégradation résiduelle). *Radiodif.-Télév.*, Vol. 4/5, **69**, 43-44.

MICELI, S. et ORLANDO, A. [octobre 1977] Sampling procedures and goodness of Fit. Proc. International Symposium on Measurement in Telecommunication (URSI), Lannion, France.

PROSSER, R. D., ALLNATT, J. W. et LEWIS, N. W. [mars 1964] Quality grading of impaired television picture. *Proc. IEE*, Vol. 111, **3**, 491-502.

SALLIO, P. et KRETZ, F. [avril-mai 1978] Qualité subjective en télévision numérique. Première partie: méthodologie de son évaluation. *Radiodif.-Télév.*, Vol. 2/5, **52**, 13-19.

WHITE, T. A. [1980] Transmission of alphanumerics by television: assessment of typescript by «experts», Proc. 9th International Symposium on Human Factors in Telecommunications, New Jersey, 27-34, Etats-Unis d'Amérique.

WHITE, T. A. [1981] Transmission of alphanumerics by television, *Displays*, **2**, 295-299.

WHITE, T. A. et REID, G. M. [août 1981] Quality of PAL colour television pictures impaired by random noise: stability of subjective assessment. *Proc. IEE*, Vol. 128, Pt F, **4**, 231-236.

Documents du CCIR

[1974-78]: 11/360 (France).

[1978-82]: 11/17 (UER); 11/259 (France); 11/287 (Royaume-Uni); 11/309 (Italie); 11/312 (France); 11/331 (République démocratique allemande); 11/357 (Italie).