Union internationale des télécommunications

UIT-R

Secteur des Radiocommunications de l'UIT

Recommandation UIT-R BS.1534-3

(10/2015)

Méthode d'évaluation subjective du niveau de qualité intermédiaire des systèmes audio

Série BS

Service de radiodiffusion sonore

ITU Union internationale des télécommunications

Avant-propos

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

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

Politique en matière de droits de propriété intellectuelle (IPR)

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

Séries des Recommandations UIT-R

(Egalement disponible en ligne: http://www.itu.int/publ/R-REC/fr)

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

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

Publication électronique
Genève, 2015

RECOMMANDATION UIT-R BS.1534-3*

Méthode d'évaluation subjective du niveau de qualité intermédiaire des systèmes audio

(2001-2003-2014-2015)

Domaine d'application

La présente Recommandation décrit une méthode d'évaluation subjective de la qualité audio intermédiaire. Cette méthode incorpore de nombreux aspects de la Recommandation UIT-R BS.1116 et fait appel à la même échelle d'évaluation que celle qui est utilisée pour l'évaluation de la qualité de l'image (voir la Recommandation UIT-R BT.500). Cette méthode, appelée «test multi stimuli avec référence et repère cachés (MUSHRA, *MUlti Stimulus test with Hidden Reference and Anchor*)», a donné de bons résultats. Les tests ont montré que la méthode MUSHRA permet d'évaluer la qualité audio intermédiaire et conduit à des résultats précis et fiables.

Mots clés

Test d'écoute, artéfacts, qualité audio intermédiaire, codage audio, évaluation subjective, qualité audio

L'Assemblée des radiocommunications de l'UIT,

considérant

a) que les Recommandations UIT-R BS.1116, UIT-R BS.1284 et UIT-R BT.500, ainsi que les Recommandations UIT-T P.800, UIT-T P.810 et UIT-T P.830, définissent des méthodes d'évaluation subjective de la qualité des systèmes audio, des systèmes vidéo et des systèmes de transmission de la parole;

b) que de nouveaux types de services de diffusion tels que les services de diffusion audio en continu sur le réseau Internet ou au moyen de lecteurs à semi-conducteurs, les services numériques par satellite, les systèmes numériques à ondes courtes ou à ondes moyennes, ou les applications multimédias pour mobiles, peuvent être exploités à un niveau de qualité audio intermédiaire;

c) que la Recommandation UIT-R BS.1116 est destinée à l'évaluation de faibles dégradations et ne convient pas pour évaluer des systèmes de qualité audio intermédiaire;

d) que la Recommandation UIT-R BS.1284 ne fournit pas de barème absolu pour l'évaluation de la qualité audio intermédiaire;

e) que l'incorporation de repères appropriés et pertinents dans les tests assure une utilisation stable de l'échelle d'évaluation subjective;

f) que les Recommandations UIT-T P.800, UIT-T P.810 et UIT-T P.830 portent essentiellement sur les signaux vocaux dans un environnement téléphonique et s'avèrent insuffisantes pour l'évaluation des signaux audio dans un environnement de radiodiffusion;

g) que l'utilisation de méthodes normalisées de test subjectif est importante pour l'échange, la compatibilité et l'évaluation correcte des résultats des tests;

h) que les nouveaux services multimédias peuvent nécessiter une évaluation conjointe des qualités audio et vidéo;

---

* La Commission d'études 6 des radiocommunications a apporté des modifications d'ordre rédactionnel à cette Recommandation en mars 2023 conformément à la Résolution 1 de l'UIT-R 1.

i) que la désignation MUSHRA est souvent utilisée à mauvais escient pour des tests n'employant pas les mêmes références et repères;

j) que les repères peuvent avoir une incidence sur les résultats des tests et qu'il est souhaitable qu'ils ressemblent aux artéfacts des systèmes testés;

k) que l'introduction de systèmes à son stéréophonique multivoies pouvant comprendre jusqu'à 3/2 voies définis dans la Recommandation UIT-R BS.775 et de systèmes sonores évolués définis dans la Recommandation UIT-R BS.2051, avec ou sans image associée, exige de nouvelles méthodes d'évaluation subjective y compris les conditions expérimentales,

recommande

1 que les procédures de test et d'évaluation figurant à l'Annexe 1 de la présente Recommandation soient utilisées pour l'évaluation subjective de la qualité audio intermédiaire,

recommande en outre

1 que les études des repères dont les caractéristiques sont celles des dégradations présentes dans les systèmes audio de pointe soient poursuivies et que la présente Recommandation soit mise à jour pour incorporer les nouveaux repères comme il convient.

Annexe 1

1 Introduction

La présente Recommandation décrit une méthode d'évaluation subjective de la qualité audio intermédiaire. Cette méthode incorpore de nombreux aspects de la Recommandation UIT-R BS.1116 et fait appel à la même échelle d'évaluation que celle qui est utilisée pour l'évaluation de la qualité de l'image (voir la Recommandation UIT-R BT.500).

Cette méthode, appelée «test multi stimuli avec référence et repère cachés (MUSHRA, *MUlti Stimulus test with Hidden Reference and Anchor*)», a donné de bons résultats. Les tests ont montré que la méthode MUSHRA permet d'évaluer la qualité audio intermédiaire et conduit à des résultats précis et fiables [2, 4, 3].

La présente Recommandation comprend les sections et annexes suivantes:

Section 1: Introduction

Section 2: Domaine d'application, objectif des tests et but de la nouvelle méthode

Section 3: Conception des expériences

Section 4: Sélection des estimateurs

Section 5: Méthode de test

Section 6: Attributs

Section 7: Séquences de test

Section 8: Conditions d'écoute

Section 9: Analyse statistique

Section 10: Rapport de test et présentation des résultats

2 Domaine d'application, objectif des tests et but de la nouvelle méthode

Les tests d'écoute subjectifs sont toujours considérés comme le moyen le plus fiable de mesure de la qualité des systèmes audio. Ils sont bien décrits et constituent des méthodes éprouvées d'évaluation de la qualité audio lorsque celle-ci est située au haut ou au bas de l'échelle.

La Recommandation UIT-R BS.1116, intitulée «Méthodes d'évaluation subjective des dégradations faibles dans les systèmes audio y compris les systèmes sonores multivoies», est employée pour évaluer les systèmes audio de haute qualité qui présentent de faibles dégradations. Il existe toutefois des applications pour lesquelles une moindre qualité des sons audio est acceptable ou inévitable. Le rapide développement de l'utilisation de l'Internet pour la diffusion et la radiodiffusion de données audio à des débits limités ont conduit à un compromis en matière de qualité audio. D'autres applications associées à une qualité audio intermédiaire sont celles qui emploient la modulation d'amplitude numérique (par exemple, le système Digital Radio Mondiale (DRM), la radiodiffusion numérique par satellite, les circuits de commentaires radio- et télédiffusés, les services audio à la demande et les lignes commutées audio). La méthode de test définie dans la Recommandation UIT-R BS.1116 ne convient pas parfaitement à l'évaluation de ces systèmes audio de moindre qualité [4], parce qu'elle ne permet pas de bien discerner les petites différences de qualité en bas de l'échelle.

La Recommandation UIT-R BS.1284 ne contient que des méthodes qui concernent la gamme audio de haute qualité ou qui ne permettent pas une évaluation absolue de la qualité audio.

D'autres Recommandations, telles que les Recommandations UIT-T P.800, UIT-T P.810 ou UIT-T P.830, traitent spécifiquement de l'évaluation subjective des signaux vocaux dans un environnement téléphonique. Le Groupe du projet B/AIM de l'Union Européenne de Radio-Télévision (UER) a effectué, à l'aide de ces méthodes de l'UIT-T, des expériences sur les séquences audio propres à un environnement de radiodiffusion. Aucune de ces méthodes ne répond aux conditions qui permettent d'obtenir une échelle absolue, de faire une comparaison avec un signal de référence ou d'avoir de petits intervalles de confiance avec un nombre raisonnable d'estimateurs. L'évaluation des signaux audio dans un environnement de radiodiffusion ne peut donc pas être effectuée correctement au moyen de l'une de ces méthodes.

La méthode de test révisée, décrite dans la présente Recommandation, vise à fournir une mesure fiable et reproductible des systèmes dont la qualité audio est normalement située dans la moitié inférieure de l'échelle de dégradation utilisée dans la Recommandation UIT-R BS.116 [2,4,3]. Dans la méthode de test MUSHRA, un signal de référence de haute qualité est utilisé et les systèmes testés devraient introduire des dégradations importantes. La méthode MUSHRA doit être employée pour l'évaluation des systèmes audio de qualité intermédiaire. Si elle est employée avec un contenu approprié, les notes

de l'auditeur seraient idéalement comprises entre 20 et 80 points MUSHRA. Si les notes pour la plupart des conditions de test sont comprises entre 80 et 100 points, les résultats du test pourraient ne pas être valables.

La répartition des notes sur un intervalle étroit peut éventuellement être due à l'intervention d'estimateurs naïfs, à l'emploi d'un contenu non décisif ou au choix d'un test ne convenant pas aux algorithmes de codage utilisés.

3 Conception des expériences

De nombreux types différents de stratégies de recherche sont employés pour recueillir des informations fiables dans un domaine d'intérêt scientifique. Il convient d'utiliser les méthodes expérimentales les plus structurées lors de l'évaluation subjective des dégradations des systèmes audio. Les expériences subjectives se caractérisent d'abord par une maîtrise réelle des conditions expérimentales, puis par la collecte et l'analyse des données statistiques émanant des auditeurs. La conception et la planification des expériences doivent être soignées afin que les facteurs non maîtrisés, pouvant créer des ambiguïtés dans les résultats des tests, soient minimisés. Par exemple, si la séquence effective des éléments sonores est la même pour tous les estimateurs qui participent à un test d'écoute, il n'est pas certain que les jugements portés par les estimateurs soient en rapport avec cette séquence plutôt qu'avec les différents niveaux de dégradation présentés. En conséquence, les conditions de test doivent être telles qu'elles mettent en évidence les effets des facteurs indépendants et d'eux seuls.

Dans les cas où l'on s'attend à ce que les dégradations possibles et les autres caractéristiques soient réparties de manière homogène tout au long du test d'écoute, l'établissement des conditions de test peut être rendu tout à fait aléatoire. Si l'on s'attend à une répartition hétérogène, il faut en tenir compte dans l'établissement des conditions de test. Par exemple, si les séquences à évaluer présentent des difficultés variables, il faut présenter les stimuli de manière aléatoire, tant au cours d'une même session qu'entre les sessions.

Les tests d'écoute doivent être conçus de manière à ne pas surcharger les estimateurs au point de rendre leur jugement moins précis. Sauf lorsque la relation entre le son et la vision est importante, il est préférable que l'évaluation des systèmes audio s'effectue sans images associées. L'intégration de conditions de contrôle appropriées est un point essentiel. Il s'agit généralement de présenter des séquences audio non endommagées, qui sont introduites de manière que les estimateurs ne puissent pas les prévoir. Les différences entre les jugements portés sur ces stimuli de contrôle et ceux qui concernent les stimuli éventuellement endommagés permettent de déterminer si les notes donnent une évaluation correcte des dégradations.

Quelques-unes de ces considérations seront abordées plus tard. Il convient de préciser que la conception des expériences, leur réalisation et l'analyse statistique sont des processus complexes, dont tous les détails ne peuvent être donnés dans une Recommandation comme celle-ci. Il est recommandé que des experts, ayant des connaissances dans le domaine de la conception des expériences et dans celui de la statistique, soient consultés ou soient présents au début de la planification du test d'écoute.

Afin d'effectuer efficacement l'analyse et le transfert des données entre laboratoires, le protocole d'expérience doit être communiqué. Tant les variables dépendantes que les variables indépendantes doivent être définies précisément. Le nombre de variables indépendantes et leurs niveaux associés doivent être définis.

4 Sélection des estimateurs

Les données obtenues au cours des tests d'écoute, qui permettent d'évaluer les faibles dégradations des systèmes audio, comme dans la Recommandation UIT-R BS.1116, doivent émaner d'estimateurs compétents en matière de détection de ces faibles dégradations. Plus la qualité des systèmes à tester est élevée, plus il est important que les auditeurs soient expérimentés.

4.1 Critères de sélection des estimateurs

Bien que la méthode de test MUSHRA ne soit pas destinée à l'évaluation des faibles dégradations, il est cependant recommandé de faire appel à des auditeurs expérimentés afin de garantir la validité des données recueillies. Ces auditeurs doivent avoir l'habitude d'écouter les sons de façon critique. Des résultats plus fiables seront ainsi obtenus plus rapidement qu'avec des auditeurs sont non expérimentés. Il est également important de noter que la plupart des auditeurs non expérimentés acquièrent en général une plus grande sensibilité aux différents types d'artéfacts après les avoir fréquemment rencontrés. Un estimateur expérimenté est choisi pour son aptitude à procéder à un test d'écoute. Cette aptitude doit être qualifiée et quantifiée en termes de ses compétences, pour ce qui est de la fiabilité et du discernement dont il fait preuve lors de la répétition des évaluations. Ceux-ci sont définis comme suit:

– **Le discernement**: L'aptitude à percevoir des différences entre les éléments testés.

– **La fiabilité**: Le fait pour les évaluations répétées du même élément testé d'être proches les unes des autres.

Seuls les estimateurs classés parmi les *estimateurs expérimentés* pour un test donné doivent participer à l'analyse finale des données. Un certain nombre de techniques d'analyse des estimateurs sont disponibles. Pour de plus amples informations, veuillez consulter le Rapport UIT-R BS.2300[1]. Ces techniques, qui sont fondées sur la répétition par chaque estimateur d'au moins une évaluation, permettent de qualifier et de quantifier sa compétence dans le cadre d'une expérience. Ces méthodes doivent être appliquées soit à la présélection des estimateurs lors d'une expérience pilote, soit de préférence à la présélection et au cours du test lui-même. Une expérience pilote est associée à une série d'expériences. Elle comporte un ensemble représentatif d'échantillons de test à évaluer dans le cadre de l'expérience principale. Aux fins de l'évaluation de la compétence de l'auditeur, l'expérience pilote doit comprendre un sous-ensemble pertinent de stimuli de test, représentatif de la gamme entière des stimuli et des artéfacts à évaluer au cours de la ou des expériences principales.

La représentation graphique de l'analyse doit contenir des informations concernant la fiabilité en fonction du discernement des estimateurs.

4.1.1 Présélection des estimateurs

Le groupe d'auditeurs doit être composé d'auditeurs expérimentés, c'est-à-dire de personnes ayant une bonne compréhension de la méthode d'évaluation subjective de la qualité et ayant reçu la formation appropriée. Ces auditeurs devraient:

– avoir l'habitude d'écouter les sons de façon critique;

– disposer de capacités auditives normales (la norme ISO 389 faisant office de référence à cet égard).

---

[1] La méthode eGauge (*expertise gauge*) comme décrite dans le Rapport UIT-R BS:2300-0 est un exemple de l'application d'une telle technique. Elle est disponible à l'adresse suivante: http://www.itu.int/oth/R0A07000036.

Le processus de formation peut aider à la présélection. Seuls les auditeurs classés comme *estimateurs expérimentés* dans le cadre d'une expérience pilote ou de l'expérience principale participent à l'analyse des données.

L'incorporation de stimuli répétés sert à l'évaluation de la fiabilité de l'auditeur.

L'argument majeur en faveur de l'introduction d'une technique de présélection réside dans le fait qu'elle permet d'accroître l'efficacité du test d'écoute. Il faut cependant veiller à ce que cette technique ne limite pas trop la validité des résultats.

4.1.2 Post-sélection des estimateurs

La méthode de post-sélection permet d'exclure comme suit les estimateurs qui attribuent une note très élevée à un signal repère sérieusement endommagé, et ceux qui notent souvent la référence cachée comme si elle était sérieusement endommagée:

– un estimateur doit être exclu de l'ensemble des réponses s'il attribue à la référence cachée, pour plus de 15% des éléments testés, une note inférieure à 90;

– un estimateur doit être exclu de l'ensemble des réponses, s'il attribue au repère de niveau moyen, pour plus de 15% des éléments testés, une note supérieure à 90. Si plus de 25% des estimateurs attribuent au repère de niveau moyen une note supérieure à 90, cela peut signifier que le repère n'a pas suffisamment été endommagé pour le test de cet élément. Dans ce cas, les estimateurs ne doivent pas être exclus sur la base des notes pour cet élément.

Cette étape initiale peut, si nécessaire, être exécutée avant que tous les estimateurs aient achevé leurs tests (de manière à permettre aux laboratoires d'essai d'estimer avant la fin des tests s'ils disposent d'un nombre suffisant d'estimateurs fiables).

Il peut être avantageux d'étudier les données afin de recenser celles qui sont erronées et aberrantes et de les soumettre à une analyse plus poussée. Une méthode appropriée consiste à employer la comparaison des notes individuelles avec toutes les notes de l'intervalle interquartile, pour une condition de test donnée *j* et une séquence audio *k*.

La médiane $\hat{x}$ et les quartiles *Q* doivent être calculés comme suit:

$$\hat{x} \coloneqq Q_2(x_{jk}) = \text{médiane}(x) \coloneqq \begin{cases} x_{jk\frac{n+1}{2}}, & n \text{ impair} \\ \frac{1}{2}\left(x_{jk\frac{n}{2}} + x_{jk\frac{n}{2}+1}\right), & n \text{ pair} \end{cases}, \; x \text{ en ordre croissant}$$

$$Q_1(x_{jk}) = \begin{cases} \text{médiane}\left(x_{jk1}, \ldots, x_{jk\frac{n+1}{2}}\right), & n \text{ impair} \\ \text{médiane}\left(x_{jk1}, \ldots, x_{jk\frac{n}{2}}\right), & n \text{ pair} \end{cases},$$

$$Q_3(x_{jk}) = \begin{cases} \text{médiane}\left(x_{jk1}, \ldots, x_{jk\frac{n+1}{2}}\right), & n \text{ impair} \\ \text{médiane}\left(x_{jk\frac{n}{2}+1}, \ldots, x_{jkn}\right), & n \text{ pair} \end{cases}.$$

L'intervalle interquartile est obtenu au moyen de l'équation $IQR(x) \coloneqq Q_3(x) - Q_1(x)$.

Dans ce contexte, les valeurs aberrantes appartiennent à l'ensemble $O(x_{jk})$:

$$O(x_{jk}) \coloneqq \{x_{jk} | x_{jk} > Q_3(x_{jk}) + 1{,}5 \cdot IQR(x_{jk})\} \cup \{x_{jk} | x_{jk} < Q_1(x_{jk}) - 1{,}5 \cdot IQR(x_{jk})\}.$$

Si, pour un stimulus particulier, une note x, attribuée par un sujet au système testé, est un élément de $O(x)$, il convient d'examiner la raison de cette notation. L'examen de l'enregistrement d'une session de test peut révéler des problèmes techniques affectant l'équipement ou une erreur humaine. L'interrogation de l'estimateur peut indiquer si la note attribuée est réellement représentative de son opinion subjective ou non. S'il est démontré que la présence d'une donnée aberrante est une erreur, elle peut être supprimée de l'ensemble des données avant l'analyse finale, le motif en étant consigné dans le rapport de test.

L'application d'une méthode de post-sélection peut rendre plus claire les tendances qui se dégagent des résultats de test. Il convient cependant de rester prudent et de garder à l'esprit les différences de sensibilité des auditeurs aux divers artéfacts. L'augmentation des effectifs du groupe d'écoute permettra de réduire l'incidence des notes individuelles d'un estimateur.

4.2 Effectifs d'un groupe d'écoute

La taille convenable d'un groupe d'écoute peut être déterminée lorsque la variance des notes attribuées par les différents estimateurs peut être estimée et que la résolution requise pour l'expérience est connue.

Lorsque les conditions d'un test d'écoute sont strictement contrôlées tant du point de vue technique que comportemental, l'expérience montre que les données d'une vingtaine d'estimateurs seulement sont souvent suffisantes pour tirer du test des conclusions appropriées. Si l'analyse peut être effectuée alors que le test est en cours, il n'est pas nécessaire d'ajouter des estimateurs dès lors que le niveau atteint en matière de signification statistique permet de formuler des conclusions appropriées.

Lorsque, pour une raison quelconque, un contrôle expérimental strict ne peut être assuré, il peut être nécessaire d'augmenter le nombre d'estimateurs pour obtenir la résolution requise.

La taille du groupe d'écoute ne dépend pas seulement de la résolution souhaitée. Le résultat du type d'expérience décrit dans la présente Recommandation n'est en principe valable que pour le groupe d'auditeurs expérimentés participant effectivement au test. En conséquence, l'augmentation de la taille du groupe d'écoute permet d'affirmer que les résultats sont valables pour un groupe d'auditeurs expérimentés plus large et qu'ils pourraient donc être considérés comme plus convaincants. Il peut être également nécessaire d'augmenter la taille du groupe d'écoute pour tenir compte de la variabilité des sensibilités des estimateurs aux différents artéfacts.

5 Méthode de test

La méthode MUSHRA fait appel au programme d'origine non modifié, la largeur de bande totale étant employée pour le signal de référence (également utilisé comme référence cachée) ainsi que pour un nombre de repères cachés obligatoires.

Des repères cachés supplémentaires peuvent être employés, de préférence ceux qui font l'objet d'autres Recommandations pertinentes de l'UIT-R. Parce que les propriétés des repères peuvent avoir un effet significatif sur les résultats d'un test, il doit être tenu compte, lors de la conception d'un repère autre qu'un repère type, des comportements optimaux des repères décrits dans l'Appendice 5. La nature de tout repère autre qu'un repère type employé dans un test doit être décrite en détail dans le rapport de test.

5.1 Description des signaux de test

Il est recommandé que la longueur maximale des séquences soit de 10 s environ, et ne dépasse pas 12 s de préférence. Ceci doit permettre d'éviter la fatigue des auditeurs, d'accroître la solidité et la stabilité de leurs réponses et de réduire la durée totale du test d'écoute. Cette durée est aussi nécessaire

pour assurer la cohérence du contenu pendant la durée entière du signal, de manière à augmenter la cohérence des réponses des auditeurs. En outre, une durée plus courte permettra aussi aux auditeurs de comparer une partie continue plus grande des signaux de test.

Si les signaux sont trop longs, les réponses des auditeurs sont influencées par des effets de prédominance et de proximité dans le temps des signaux de test, ou par des zones isolées en boucle dont les caractéristiques spectrales et temporelles peuvent varier beaucoup pendant la durée su signal. Le raccourcissement de la durée des signaux de test doit permettre de réduire ces différences. Mais cette limitation peut ne pas convenir dans certaines circonstances. Un exemple peut être celui d'un test portant sur un son variant lentement au cours d'une longue période. Dans ces conditions limites, dans lesquelles il est indiqué d'employer un stimulus plus long, il est nécessaire d'étayer et de justifier dans le rapport de test final pourquoi il a fallu accroître la durée.

L'ensemble des signaux transformés comprend tous les signaux testés ainsi qu'au moins deux signaux «repères» supplémentaires. Le repère type est un signal obtenu après filtrage passe-bas à une fréquence de coupure de 3,5 kHz, le repère de qualité moyenne ayant une fréquence de coupure de 7 kHz.

Les largeurs de bande des repères correspondent à celles des Recommandations relatives aux circuits de contrôle (3,5 kHz), utilisés à des fins de surveillance et de coordination en radiodiffusion, aux circuits de commentaires (7 kHz) et aux circuits occasionnels (10 kHz), conformément aux Recommandations UIT-T G.711, UIT-T G.712, UIT-T G.722 et UIT-T J.21, respectivement.

Les caractéristiques du filtre passe-bas à 3,5 kHz doivent être les suivantes:

f_c = 3,5 kHz

Ondulation maximale dans la bande-passante = ±0,1 dB

Affaiblissement minimal à 4 kHz = 25 dB

Affaiblissement minimal à 4,5 kHz = 50 dB.

Les repères supplémentaires sont prévus pour permettre une comparaison des systèmes testés avec des niveaux de qualité audio bien connus. Ils ne doivent pas être utilisés pour réévaluer les résultats entre différents tests.

5.2 Phase de formation

Afin d'aboutir à des résultats fiables, il est nécessaire de former les estimateurs lors de sessions spéciales préalables au test. Une telle formation s'est révélée importante pour obtenir des résultats fiables. La formation devrait au moins mettre le sujet en présence de toute la gamme et de tous les types de dégradations, ainsi que de l'ensemble des signaux employés au cours du test. Plusieurs méthodes peuvent être utilisées à cet effet: un simple système de lecture à bande magnétique ou un système interactif assisté par ordinateur. Des instructions figurent à l'Appendice 1. La formation devrait aussi servir à faire en sorte que les estimateurs se familiarisent avec les outils de tests subjectifs (par exemple, les logiciels de test).

5.3 Présentation des stimuli

La méthode MUSHRA est une méthode de test en double aveugle à stimuli multiples avec référence cachée et repères cachés, alors que la méthode de la Recommandation UIT-R BS.1116 est une méthode de test «en double aveugle à triple stimuli avec référence cachée». Il est considéré que la méthode MUSHRA convient mieux à l'évaluation des dégradations moyennes et fortes [4].

Lors d'un test impliquant de faibles dégradations, la difficulté pour le sujet est de détecter tous les artéfacts susceptibles d'être présents dans le signal. Il est nécessaire dans ce cas de disposer d'un signal de référence cachée pour permettre à l'expérimentateur d'évaluer l'aptitude du sujet à bien détecter ces artéfacts. En revanche, dans un test concernant des dégradations moyennes et fortes, le sujet n'éprouve aucune difficulté à détecter les artéfacts, et une référence cachée est donc inutile dans ce cas. La difficulté réside plutôt dans le fait que le sujet doit noter les nuisances relatives des différents artéfacts. Il doit alors pondérer ses préférences pour l'un ou l'autre type d'artéfact.

L'utilisation d'une référence de haute qualité conduit à un problème intéressant. Puisque cette nouvelle méthode doit être utilisée pour évaluer des dégradations moyennes et fortes, l'on s'attend à ce que la différence perçue entre le signal de référence et les éléments de test soit relativement grande. En revanche, la différence perçue entre les éléments de test qui appartiennent à des systèmes différents peut être très faible. En conséquence, si une méthode de test à essais multiples (comme celle qui est utilisée dans la Recommandation UIT-R BS.1116) est employée, il peut être très difficile pour les estimateurs de différencier avec précision les divers signaux dégradés. Par exemple, dans un test de comparaison directe par paires, les estimateurs peuvent s'accorder pour dire que le système A est meilleur que le système B. Cependant, dans le cas où chaque système est comparé seulement au signal de référence (les systèmes A et B ne sont pas directement comparés entre eux), les différences entre les deux systèmes peuvent disparaître.

Afin de surmonter cette difficulté dans la méthode de test MUSHRA, le sujet peut à sa guise passer du signal de référence à l'un quelconque des systèmes testés, généralement en utilisant un système de lecture assisté par ordinateur, bien que d'autres mécanismes employant des lecteurs de CD multiples ou des magnétophones à bandes multiples peuvent être employés. Le sujet se voit présenter une suite d'essais. Pour chaque essai, on lui soumet la version de référence, le repère de faible et de moyenne qualité ainsi que toutes les versions du signal de test transformées par les systèmes testés. Par exemple, si un test comporte 8 systèmes audio, le sujet peut presque instantanément passer des 11 signaux de test à un signal de référence ouverte (1 référence + 8 systèmes testés + 1 référence cachée + 1 repère caché de faible qualité + 1 repère caché de moyenne qualité).

Puisque le sujet peut directement comparer les signaux dégradés, cette méthode offre les avantages d'un test de comparaison entièrement effectuée par paires, le sujet pouvant plus facilement détecter les différences entre les signaux dégradés et les noter en conséquence. Cette caractéristique permet d'obtenir une résolution élevée pour les notes attribuées aux systèmes. Il est important de noter toutefois que les estimateurs détermineront leur note pour un système donné après avoir comparé ce système au signal de référence, ainsi qu'aux autres signaux de chacun des essais.

Il est recommandé de ne pas faire intervenir plus de 12 signaux par essai (par exemple, 9 systèmes testés, 1 repère caché de faible qualité, 1 repère caché de moyenne qualité et 1 référence cachée).

Dans les rares cas où un grand nombre de signaux doivent être comparés, il peut être nécessaire d'employer une structure en blocs de l'expérience, qui sera consignée en détail.

Dans les tests de la Recommandation UIT-R BS.1116, les estimateurs ont tendance à aborder un essai donné en commençant par un processus de détection, suivi d'un processus de notation. L'expérience de la conduite des tests selon la méthode MUSHRA montre que les estimateurs ont tendance à commencer une session par une estimation grossière de la qualité. Ils procèdent ensuite au tri ou au classement, puis à la notation. Puisque le classement est effectué directement, les résultats concernant la qualité audio intermédiaire sont susceptibles d'être plus cohérents et fiables que lorsque la méthode de la Recommandation UIT-R BS.1116 est employée. En outre, la durée minimale de la boucle est de 500 ms et une augmentation et une diminution progressives pendant 5 ms de l'enveloppe en cosinus surélevé doivent être appliquées à tout contenu en boucle. Tout passage en matière de contenu d'un système testé à un autre doit comporter une augmentation et une diminution progressives pendant 5 ms de l'enveloppe en cosinus surélevé. A aucun moment au cours du test, il ne faut employer

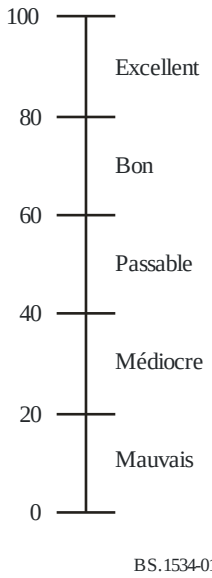
d'enchaînement avec chevauchement lors du passage d'un système testé à un autre. Ces modifications permettent de réduire l'emploi des changements de la coloration spectrale au cours des comparaisons soudaines de courte durée visant à identifier et à noter les signaux testés.

5.4 Procédure de notation

Les estimateurs doivent noter les stimuli suivant l'échelle de qualité continue (CQS). L'échelle CQS comporte des échelles graphiques identiques (généralement d'une longueur d'au moins 10 cm), divisées en cinq intervalles égaux repérés par les adjectifs indiqués de haut en bas dans la Fig. 1.

Cette échelle est également utilisée pour l'évaluation de la qualité de l'image (Recommandation UIT-R BT.500 – Méthodologie d'évaluation subjective de la qualité des images de télévision).

FIGURE 1



BS.1534-01

L'auditeur enregistre son évaluation sous une forme qui convient, par exemple, en employant des curseurs sur un dispositif d'affichage électronique (voir la Fig. 2), ou en utilisant un crayon et une graduation sur papier. Dans le cas d'un dispositif semblable à celui de la Fig. 2, il faut faire en sorte que le sujet soit en mesure de modifier uniquement la note qu'il a attribuée à l'élément en cours d'écoute. Quelques orientations en matière de conception des interfaces sont données à l'Appendice 2. L'estimateur est prié de noter la qualité de tous les stimuli suivant l'échelle CQS à cinq intervalles.

FIGURE 2

Exemple d'un affichage sur ordinateur utilisé pour un test MUSHRA

BS.1534-02

La méthode MUSHRA présente l'avantage, par rapport à celle de la Recommandation UIT-R BS.1116, d'afficher de nombreux stimuli en même temps, de manière que le sujet soit en mesure de les comparer directement entre eux. La durée nécessaire pour exécuter le test à l'aide de la méthode MUSHRA est sensiblement plus courte que celle de la méthode de la Recommandation UIT-R BS.1116.

5.5 Enregistrement des sessions de test

Si une anomalie est observée lors du traitement des notes attribuées, il est très utile de disposer d'un enregistrement des événements sur lesquels portent ces notes. Pour ce faire, un moyen relativement simple consiste à effectuer des enregistrements vidéo et audio de l'ensemble du test. Dans le cas où une note anormale est repérée dans un ensemble de résultats, l'enregistrement sur bande magnétique peut être examiné afin de tenter d'établir si la cause en est une erreur humaine ou un dysfonctionnement de l'équipement.

6 Attributs

Les attributs propres aux évaluations monophoniques, stéréophoniques et multivoies sont énumérés ci-après. Il est préférable d'évaluer l'attribut «qualité audio de base» dans chaque cas. Les expérimentateurs peuvent choisir de définir et d'évaluer d'autres attributs.

Il convient de ne noter qu'un seul attribut par essai. S'il est demandé aux estimateurs d'évaluer plus d'un attribut par essai, ils peuvent se sentir débordés ou embrouillés, ou les deux, en tentant de répondre à de multiples questions concernant un stimulus donné. Cela peut conduire à une notation peu fiable pour l'ensemble des questions. Si plusieurs propriétés du signal audio doivent être évaluées indépendamment, il est recommandé d'évaluer d'abord la qualité audio de base.

6.1 Système monophonique

Qualité audio de base: cet attribut unique et global est utilisé pour évaluer toutes les différences détectées entre la référence et l'objet.

6.2 Système stéréophonique

Qualité audio de base: cet attribut unique et global est utilisé pour évaluer toutes les différences détectées entre la référence et l'objet.

L'attribut supplémentaire suivant peut présenter de l'intérêt:

Qualité de l'image stéréophonique: cet attribut est associé aux différences entre la référence et l'objet en termes d'emplacement des images sonores et de sensations de profondeur et de réalité de l'événement audio. Bien que certaines études aient montré que la qualité de l'image stéréophonique peut être dégradée, cette question n'a pas encore fait l'objet de recherches suffisantes pour justifier une notation de la qualité de l'image stéréophonique, distincte de celle de la qualité audio de base.

NOTE 1 – Jusqu'en 1993, la plupart des études de l'évaluation subjective des faibles dégradations, qui concernent les systèmes stéréophoniques, ne faisaient intervenir que la qualité audio de base. L'attribut qualité de l'image stéréophonique était donc, dans ces études, implicitement ou explicitement intégré dans la qualité audio de base, en tant qu'attribut global.

6.3 Système multivoies

Qualité audio de base: cet attribut unique et global est utilisé pour évaluer toutes les différences détectées entre la référence et l'objet.

Les attributs supplémentaires suivants peuvent présenter de l'intérêt:

Qualité frontale de l'image: cet attribut est associé à la localisation des sources sonores frontales. Il porte sur la qualité de l'image stéréophonique et les pertes de définition.

Qualité de la sensation ambiophonique: cet attribut est associé à une sensation spatiale, à l'ambiance ou à des effets ambiophoniques directionnels particuliers.

6.4 Système sonores évolués

Qualité audio de base: On utilise cette caractéristique unique et globale pour évaluer toutes les différences décelées entre la référence et l'objet de l'essai. L'examen de cette caractéristique pour les systèmes sonores évolués devrait également porter sur toutes les caractéristiques décrites pour les systèmes multivoies.

En outre, les caractéristiques ci-après peuvent présenter un intérêt:

Qualité du timbre: Il a été établi que cette caractéristique est très importante. La caractéristique de qualité du timbre peut être décrite par deux ensembles de propriétés:

– Le premier ensemble de propriétés relatives au timbre concerne la *couleur sonore*, par exemple l'éclat, le timbre musical, la coloration, la clarté, la dureté, l'égalisation ou la richesse.

– Le second ensemble de propriétés relatives au timbre concerne *l'homogénéité sonore*, par exemple la stabilité, la pureté, le réalisme, la fidélité et les nuances. Ces propriétés peuvent permettre de décrire le timbre sonore, mais également d'autres caractéristiques sonores.

Qualité de localisation: Cette caractéristique correspond à la localisation de toutes les sources sonores directionnelles. Elle comprend la qualité de l'image stéréophonique et les pertes de définition. Cette caractéristique peut être divisée en *qualité de localisation horizontale*, en *qualité de localisation verticale* et en *qualité de localisation distante*. Si les essais sont accompagnés d'images, ces caractéristiques peuvent également être divisées en *qualité de localisation sur l'écran* et en *qualité de localisation autour de l'auditeur*.

Qualité de l'environnement: Cette caractéristique est une extension de la caractéristique qualité ambiophonique. Cette caractéristique correspond à l'impression d'espace, à l'enveloppement, à l'ambiance, à la diffusivité ou aux effets spatiaux directionnels d'immersion. Elle peut être divisée en *qualité de l'environnement horizontale*, en *qualité de l'environnement verticale* et en *qualité de l'environnement distante*.

7 Séquences de test

7.1 Elément de test

Il convient d'utiliser des séquences décisives, représentatives d'un programme radiodiffusé type pour l'application souhaitée, qui permettent de faire apparaître les différences entre les systèmes testés. Les séquences sont décisives si elles sont contraignantes pour les systèmes testés. Il n'existe pas de programme susceptible de convenir partout, qui puisse être utilisé pour évaluer tous les systèmes dans toutes les conditions. En conséquence, pour chaque système testé et chaque expérience il faut rechercher des séquences de programme décisives qui conviennent. La recherche de séquences adaptées prend en général beaucoup de temps. Cependant, à moins de trouver des séquences réellement décisives pour chaque système, les expériences ne permettront pas de faire apparaître des différences entre les systèmes, ni de dégager des conclusions. Un petit groupe d'auditeurs expérimentés doit choisir les éléments à tester parmi une sélection plus large d'éléments admissibles possibles. Ce processus de sélection doit inclure tous les systèmes à tester et être étayé et consigné dans le récapitulatif des tests.

Il doit empiriquement et statistiquement être prouvé que toute défaillance en matière de détermination des différences entre les systèmes ne provient pas d'un manque de sensibilité expérimentale, peut-être dû au mauvais choix des séquences audio, ou à un autre point faible de l'expérience. À défaut, cette détermination «nulle» ne peut être acceptée comme valable.

Dans cette recherche de séquences décisives, il faut admettre tout stimulus qui peut être considéré comme une séquence radiodiffusée possible. Des signaux synthétisés, délibérément conçus pour perturber un système particulier, ne sont pas admissibles. Le contenu artistique ou intellectuel d'une suite de programmes ne doit jamais être ni tellement attractif, ni tellement désagréable ou ennuyeux qu'il détourne l'attention du sujet et l'empêche de se concentrer sur la détection des dégradations. Il convient de tenir compte de la fréquence d'apparition attendue de chaque type de séquence dans des programmes effectivement radiodiffusés. Il convient aussi de préciser que la nature des séquences radiodiffusées pourrait varier au cours du temps en fonction de l'évolution des préférences et des styles musicaux.

Lors du choix des séquences d'un programme, il est important de définir avec précision les attributs à évaluer. La responsabilité du choix des séquences doit être confiée à un groupe d'estimateurs qualifiés qui ont des connaissances de base des dégradations attendues. Leur point de départ doit être fondé sur une très large gamme de séquences, pouvant s'étendre à des enregistrements conçus pour l'occasion.

Dans le but de préparer le test subjectif proprement dit, le groupe d'estimateurs qualifiés doit régler de manière subjective le volume sonore de chaque extrait avant de l'enregistrer sur le support du test. Cela permettra par la suite d'employer ce support, dont le gain est fixé, pour tous les éléments du programme qui font l'objet d'un essai expérimental.

Pour toutes les séquences de test, le groupe d'estimateurs qualifiés doit se réunir et se mettre d'accord sur les niveaux sonores relatifs des différents extraits de test. En outre, pour la séquence dans son ensemble, les experts doivent s'accorder sur le niveau absolu de pression sonore reproduit par rapport au niveau d'alignement.

Une salve de tonalité (par exemple 1 kHz, 300 ms, –18 dBFS) au niveau du signal d'alignement peut être insérée en tête de chaque enregistrement pour permettre de régler son niveau d'alignement à la sortie par rapport au niveau d'alignement à l'entrée, requis par le canal de reproduction, conformément à la Recommandation R.68 de l'UER (voir la Recommandation UIT-R BS.1116, § 8.4.1). La salve de tonalité ne sert qu'à l'alignement: elle ne doit pas être rediffusée au cours du test. Il faut contrôler le signal radiophonique de façon que les amplitudes des crêtes ne dépassent que rarement l'amplitude de crête du signal maximal admis, tel qu'il est défini dans la Recommandation UIT-R BS.645 (une onde sinusoïdale dépassant de 9 dB le niveau d'alignement).

Le nombre d'extraits que l'on peut inclure dans un test varie: il doit être le même pour tous les systèmes testés. Il semble raisonnable de le choisir égal à 1,5 fois le nombre de systèmes testés, à condition d'atteindre la valeur minimale de 5 extraits. En raison de la complexité de la tâche, il faut mettre les systèmes à tester à disposition de l'expérimentateur. Une bonne sélection ne peut être réalisée que si un calendrier approprié est défini. En outre, en raison de la variation au cours du temps de l'utilisation du débit binaire, il est recommandé de coder des séquences plus longues et d'utiliser une partie de chaque séquence dans le test d'écoute.

La qualité de fonctionnement d'un système multivoies dans des conditions de lecture à deux voies doit être testée au moyen d'un mixage réducteur de référence. Bien que le recours à un mixage réducteur fixe puisse être jugé restrictif dans certains cas, il s'agit sans conteste de l'option la plus raisonnable à utiliser à long terme par les radiodiffuseurs. Les équations relatives au mixage réducteur de référence sont les suivantes (voir la Recommandation UIT-R BS.775):

$$L_0 = 1{,}00L + 0{,}71C + 0{,}71L_S$$

$$R_0 = 1{,}00R + 0{,}71C + 0{,}71R_S$$

La présélection d'extraits de test convenant à l'évaluation critique de la qualité de fonctionnement d'un mixage réducteur de référence à deux voies doit être fondée sur la reproduction des séquences de programme transformées par mixage réducteur à deux voies.

7.2 Configuration des haut-parleurs

Au cas où une configuration des haut-parleurs différente de celle définie dans la Recommandation UIT-R BS.775 serait utilisée pour l'expérience, pour préciser les conditions d'essai, la position de tous les haut-parleurs (distances et angles) pendant les essais, ainsi que la disposition de ces haut-parleurs par rapport à la position d'écoute, doivent être décrites en détail dans le rapport d'essai. La forme et le contenu de cette description doivent correspondre aux dispositions des haut-parleurs et aux positions d'écoute définies dans la Recommandation UIT-R BS.775. Il faudra en outre indiquer et décrire la position de chaque haut-parleur dans le plan vertical pour les systèmes sonores évolués comprenant des haut-parleurs installés à des hauteurs différentes. La Recommandation UIT-R BS.2051 donne des informations pouvant être utiles dans ce contexte.

8 Conditions d'écoute

Des méthodes d'évaluation subjective des faibles dégradations dans les systèmes audio, y compris les systèmes sonores multivoies, sont définies dans la Recommandation UIT-R BS.1116. Les conditions d'écoute décrites dans les § 7 et 8 de la Recommandation UIT-R BS.1116 doivent être utilisées pour évaluer les systèmes audio de qualité intermédiaire.

Des casques d'écoute ou des haut-parleurs peuvent être employés au cours du test. L'utilisation des deux moyens lors d'une même session de test n'est pas permise: tous les estimateurs doivent utiliser le même type de transducteur.

Dans le cas d'un signal de mesure dont la tension efficace est égale au «niveau du signal d'alignement» (0 dBu0s conformément à la Recommandation UIT-R BS.645; –18 dB en dessous du niveau de coupure d'un magnétophone numérique, conformément à la Recommandation UER R.68) appliqué successivement à l'entrée de chaque voie de reproduction (par exemple un amplificateur de puissance et son haut-parleur associé), il faut régler le gain de l'amplificateur au niveau de la pression acoustique de référence (pondération CEI/A, lente) suivant:

$$L_{ref} = 78 \pm 0{,}25 \text{ dBA}$$

Le réglage individuel par un sujet du niveau d'écoute est admis en cours de session, mais dans la limite d'un intervalle de ±4 dB par rapport au niveau de référence défini dans la Recommandation UIT-R BS.1116. Le groupe de sélection doit trouver un équilibre entre les éléments d'un test qui soit tel que les estimateurs n'aient normalement pas besoin de procéder à des réglages individuels pour chaque élément.

Les réglages de niveau au sein d'un élément ne doivent pas être admis.

9 Analyse statistique

Les valeurs de la longueur sur les feuilles de notation, relatives aux évaluations pour chacune des conditions de test, sont converties linéairement en notes normalisées comprises entre 0 et 100, où 0 correspond au bas de l'échelle (mauvaise qualité). Les notes absolues sont ensuite calculées comme suit.

Une analyse statistique, soit paramétrique soit non paramétrique, peut être effectuée, si les hypothèses statistiques sont satisfaites (voir le § 9.3.3). Des orientations en matière d'analyse statistique paramétrique sont données à l'Appendice 4.

9.1 Visualisation et analyse exploratoire des données

L'analyse statistique doit toujours débuter par une visualisation des données brutes. Cela peut se faire au moyen d'histogrammes avec une courbe d'ajustement pour une distribution normale, de diagrammes en boîtes (à moustaches) ou de diagrammes quartile par quartile (Q-Q)

La visualisation des données sous la forme de diagrammes en boîtes donne des indications sur l'existence de données aberrantes et sur leurs effets dans les résumés descriptifs des données. Cette visualisation permet de déterminer la dispersion et l'écart des notes individuelles par rapport à la note médiane de l'ensemble des estimateurs. La visualisation sous la forme d'un histogramme permet de repérer la présence d'une distribution multimodale sous-jacente. Si une distribution multimodale est observée avec certitude pour les données, il est conseillé à l'expérimentateur d'analyser la distribution séparément.

Afin d'évaluer le degré de multimodalité *b*, il convient d'employer la formule suivante:

$$b = \frac{g^2 + 1}{k + \frac{3(n-1)^2}{(n-2)(n-3)}}$$

où:

n: la taille de l'échantillon

g: l'asymétrie de l'échantillon de taille finie

k: l'excès d'aplatissement des résultats du test d'écoute.

Ce coefficient est compris entre 0 et 1. Les valeurs élevées (> 5/9) indiquent la multimodalité.

Sur la base de l'examen visuel de ces diagrammes, du degré *b* et des hypothèses concernant la population sous-jacente de l'échantillon observé, il convient de décider s'il peut être supposé qu'une distribution normale a été observée ou pas. Si la courbe d'ajustement est clairement asymétrique, si l'histogramme contient de nombreuses données aberrantes ou si le diagramme Q-Q n'est pas du tout une ligne droite, il convient de ne pas considérer l'échantillon comme étant normalement distribué. Le calcul de la médiane des notes normalisées de tous les auditeurs qui ont passé la post-sélection conduira aux notes médianes subjectives.

La médiane doit être calculée comme suit: $\hat{x} = \text{médiane}(x) = \begin{cases} x_{\frac{n+1}{2}} & n \text{ impair} \\ \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) & n \text{ pair} \end{cases}$, *x* en ordre croissant.

La première étape de l'analyse correspond au calcul de la note médiane $\overline{\eta}_{jk}$ pour chacune des présentations. Il s'ensuit que η_{ijk} est la note médiane de l'observateur *i* pour une condition d'essai *j* et une séquence audio *k*, et que $\hat{\eta}$ est la médiane de l'échantillon (pour l'ensemble des observateurs, des conditions et des séquences audio).

Le calcul des notes médianes globales, $\overline{\eta}_j$ et $\overline{\eta}_k$, peut se faire de la même manière pour chacune des conditions de test et chacune des séquences de test.

Bien que l'utilisation des valeurs moyennes soit nécessaire pour certaines méthodes d'analyse telles que l'analyse de la variance (ANOVA) (voir le § 9.3), le calcul de la médiane est une autre mesure possible de la tendance centrale. La médiane est une mesure robuste de la tendance centrale, qui est optimale dans les cas où l'ensemble des échantillons est petit, la distribution est non normale ou l'ensemble des données contient des données fortement aberrantes. Il est possible que, dans beaucoup de scénarios de test, ces préoccupations se justifient moins. Toutefois, étant donné que l'un des plus grands avantages du test normalisé est qu'il permet de comparer et d'interpréter les notes pour l'ensemble des utilisateurs et des sites, il est utile de recenser les méthodes d'analyse qui sont les plus robustes et les moins sensibles aux facteurs susceptibles d'altérer la validité ou de réduire la transcription d'un test à l'autre.

De telles méthodes sont celles des statistiques non paramétriques. Lorsqu'une analyse non paramétrique des données est effectuée, les moyennes et les intervalles de confiance à 95% doivent être calculés à partir des méthodes disponibles telles que celles qui sont fondées sur un algorithme de «bootstrap» courant.

Les valeurs de l'erreur par rapport à la médiane peuvent être calculées à partir de l'écart absolu moyen:

$$\hat{\tau} = \Sigma \left| Y_i - \hat{\eta} \right| / n$$

L'intervalle interquartile (IQR) est recommandé comme mesure de la confiance autour de la médiane. C'est la différence entre les 1er et 3ème quartiles: $IQR = Q_3 - Q_1$. Les formules sont données au § 4.1.2. Si la distribution des résultats est normale, l'intervalle IQR est égal à deux fois l'écart absolu moyen.

Il est recommandé que la signification statistique soit déterminée pour un niveau de confiance à 95%. Les tests non paramétriques de simulation de distribution aléatoire sont des mesures robustes de la signification statistique. À la différence des analyses statistiques paramétriques, ils ne font aucune hypothèse concernant la distribution sous-jacente des données et sont moins sensibles aux nombreux problèmes associés à l'emploi d'une taille d'échantillon plus petite.

Un test non paramétrique robuste de simulation de distribution aléatoire (test de permutation) permet de déterminer la probabilité qu'une différence entre deux conditions de test puisse être observée si les données sont réellement aléatoires comme le suppose l'hypothèse nulle. La probabilité mesurée dans ce test est une valeur réelle, déterminée à partir de la distribution des données effectives, plutôt qu'une valeur déduite en supposant une forme donnée pour la distribution sous-jacente [5]. Cette façon d'exécuter le test nécessite des techniques courantes de rééchantillonnage telles que les techniques de «bootstrap» ou de simulation de Monte-Carlo, qui sont maintenant facilement accessibles grâce à la grande rapidité des moyens de calcul moderne [6]. Cette méthode de test est plus amplement décrite à l'Appendice 3.

9.2 Analyse de la puissance

L'analyse de la puissance peut être utile pour estimer les tailles des échantillons nécessaires aux tests d'écoute, si elle est utilisée comme analyse *a priori*, et pour estimer la puissance ou l'erreur de type II du test, si elle est employée comme analyse *a posteriori*. L'analyse *a priori* fournit la taille de l'échantillon nécessaire à l'expérience, la dimension de l'effet étant $d = \frac{\bar{x}}{s}$, le niveau de signification α et la puissance statistique $1 - \beta$.

L'analyse *a posteriori* au contraire fournit la puissance 1 – β ou l'erreur de type II β du test, la dimension de l'effet étant $d = \frac{\bar{x}}{s}$, le niveau de signification α et la taille de l'échantillon *N*. L'erreur de type II β est la probabilité que l'effet *d* existe dans la population mais n'a pas été détecté comme étant significatif lors du test. Si, par exemple, un test indique que la qualité n'est pas altérée par le système, 1 – β est la probabilité que la dégradation a été démontrée lors du test[2].

[2] De nombreux outils tels que G*Power [16] permettent d'effectuer l'analyse de la puissance automatiquement pour des distributions de population connues, mais il est plus difficile d'estimer la puissance pour des populations non connues.

9.3 Application et utilisation de l'analyse de la variance

9.3.1 Introduction

Cette section est axée sur les spécifications nécessaires à l'étude statistique paramétrique au moyen de l'analyse de la variance (ANOVA). En raison de la robustesse du modèle ANOVA (voir les références [7], [8], [12] et [13]) et de sa puissance statistique[3], cette méthodologie est bien adaptée aux données recueillies à l'aide de la méthodologie de la présente Recommandation UIT-R BS.1534. Puisque l'étude statistique fondée sur l'analyse ANOVA-F est assez robuste, s'agissant tant des distributions non normales de données que de l'hétérogénéité de la variance, le test des hypothèses est centré autour de la nature de l'erreur ou des résidus.

Pour de plus amples informations sur les hypothèses générales associées aux statistiques paramétriques, veuillez vous reporter à l'Appendice 4.

9.3.2 Spécification d'un modèle

Il est fortement conseillé qu'au cours de la conception de l'expérience (voir le § 3), le modèle soit spécifié en profondeur en termes de variables indépendantes (par exemple, ECHANTILLON, SYSTEME, CONDITION, etc.) et de variables dépendantes (par exemple, qualité audio de base ou effort d'écoute, etc.). Les niveaux de chacune des variables indépendantes doit être défini au cours de la phase de spécification du modèle.

Au cours de la définition d'un modèle d'analyse (employant, par exemple, l'analyse de la variance ANOVA ou l'analyse ANOVA à mesures répétées (rmANOVA)), il est important d'incorporer toutes les variables significatives. L'omission de variables significatives, par exemple, les interactions doubles ou triples de facteurs indépendants, peut conduire à une mauvaise spécification du modèle, qui à son tour peut conduire à une faible variance expliquée (R^2) et à une mauvaise interprétation possible de l'analyse des données.

9.3.3 Aide-mémoire destiné à l'analyse statistique paramétrique

Cet aide-mémoire donne des directives succinctes à appliquer lors de l'examen des données, la vérification des hypothèses de base (paramétriques et non paramétriques), ainsi que les étapes de base de la statistique paramétrique. L'accent est mis sur les règles à suivre pour l'analyse de la variance, en tant que méthode convenant à l'analyse des données qui proviennent des expériences décrites dans la présente Recommandation UIT-R BS.1534. Pour un guide complet, veuillez vous reporter aux manuels de statistique (par exemple, les références [8], [11] et [9]).

- Statistique exploratoire[4]
  - Examiner si la structure des données est correcte et telle que prévue
  - Vérifier si des données manquent
  - Etudier la normalité des distributions de données
  - Examiner d'autres distributions de données possibles (unimodale, bimodale, asymétrique, etc.)

---

[3] Il est généralement conseillé de choisir la méthode d'analyse statistique la plus puissante que permettent les données ([9] et [10]).

[4] Cela s'applique aussi bien à la statistique paramétrique que non paramétrique.

- Unidimensionnalité
  - Vérifier que les estimateurs emploient la même échelle[5]
  - Vérifier que les données sont unidimensionnelles
  - Analyser les principaux composants, le diagramme de Tucker-1 ou le coefficient alpha de Cronbach
- Indépendance des observations
  - Cette propriété fait habituellement partie de la méthodologie expérimentale. Elle ne peut être vérifiée facilement du point de vue de la statistique. Il faut s'assurer que les données ont été recueillies indépendamment, c'est-à-dire en employant des techniques expérimentales en double aveugle et en faisant en sorte que les estimateurs ne s'influencent pas mutuellement
- Homogénéité de la variance[6]
  - Vérifier l'hypothèse selon laquelle les variances des différentes variables indépendantes sont semblable
    - Examiner visuellement les diagrammes en boîtes placés côte à côte pour chacun des niveaux des variables indépendantes. On appliquera la règle empirique selon laquelle l'hétérogénéité varie au maximum d'un facteur 4
    - Le test de Brown et Forsythe ou la statistique de Levene peuvent être employés pour évaluer l'homogénéité de la variance
- Distribution normale des résidus
  - Vérifier la distribution normale des résidus
    - Exécuter le test D de Kolmogorov et Smirnov, le test de K-S Lillefors ou le test de Levene
    - Les diagrammes de probabilité normale (parfois nommés diagrammes P-P) ou les diagrammes quantile par quantile (souvent nommés diagrammes Q-Q) peuvent aussi être employés en tant que tests visuels de la distribution normale
- Détection des données aberrantes
  - Les données aberrantes doivent être recherchées et peut-être éliminées si cela se justifie. Des orientations concernant cette question sont données au § 4.1.2
- Analyse
  - Analyse de la variance (ANOVA) – Modèle linéaire général (GLM) ou modèle ANOVA à mesures répétées
    - Utiliser un modèle ANOVA adapté, par exemple, le modèle linéaire général (GLM) ou le modèle ANOVA à mesures répétées. De plus amples détails sont donnés à l'Appendice 4
    - Spécifier le modèle selon la conception de l'expérience
      - Incorporer si possible les interactions doubles ou triples
    - Analyser les données à l'aide du modèle ainsi que les résultats

---

[5] La multidimensionnalité a été observée dans les cas où des sous-populations avaient des opinions différentes sur l'évaluation d'artéfacts particuliers.

[6] Cette caractéristique est requise pour l'analyse ANOVA mais pas pour l'analyse rmANOVA (voir l'Appendice 4).

- Examiner la variance expliquée (R2) du modèle employé pour décrire la variable dépendante
- Examiner la distribution des erreurs résiduelles
- Examiner les facteurs significatifs et non significatifs

– Recommencer l'analyse avec le modèle pour éliminer les données aberrantes et les facteurs non significatifs

– Tests a posteriori
  – Exécuter des tests *a posteriori* afin d'établir la signification de la différence entre les moyennes, dans les cas où le facteur dépendant (ou l'interaction des facteurs) est significatif dans l'analyse ANOVA
  – Un certain nombre de différents tests *a posteriori* avec différents niveaux de discernement sont disponibles, par exemple, le test de la différence la moins significative (LSD) de Fisher, le test de la différence honnêtement significative (HSD) de Tukey, etc.
  – Il est recommandé de consigner la grandeur des effets ainsi que les niveaux de signification
– Établir des conclusions
  – Une fois l'analyse effectuée, résumer les conclusions en portant sur un diagramme les moyennes et les intervalles de confiance à 95% associés pour les données brutes ou les données modélisées selon l'analyse ANOVA (parfois nommées moyennes marginales estimées)
  – Dans les cas où les interactions (par exemple, doubles ou triples) des facteurs s'avèrent être significatives, il convient de les représenter graphiquement pour donner un aperçu complet des données. Dans ces cas, la représentation graphique des effets principaux seulement donnera un aperçu des données faussé par les effets d'interaction.

D'autres orientations en matière d'utilisation des modèles ANOVA sont données à l'Appendice 4 et dans des articles statistiques et appliqués courants, par exemple les références [11], [13] et [15].

10 Rapport de test et présentation des résultats

10.1 Observations générales

Les résultats, de par leur présentation, doivent être faciles à exploiter de manière que tout lecteur, qu'il soit novice ou expérimenté, soit en mesure d'obtenir des informations pertinentes. Tout lecteur souhaite d'abord visualiser l'ensemble des résultats de l'expérience, de préférence sous une forme graphique. Cette présentation peut s'accompagner d'informations quantitatives plus détaillées, même si les analyses numériques détaillées complètes doivent être données en appendice.

10.2 Teneur du rapport de test

Le rapport de test doit indiquer aussi clairement que possible les raisons de l'étude, les méthodes utilisées et les conclusions tirées. Suffisamment de détails doivent être donnés pour qu'une personne compétente puisse en principe reproduire cette étude afin d'en vérifier empiriquement les résultats.

Toutefois, il n'est pas nécessaire que le rapport contienne tous les résultats. Un lecteur averti doit être à même de comprendre et de formuler une critique sur les points les plus importants du test, tels que les motifs à l'origine de l'étude, les méthodes de conception et l'exécution de l'expérience, ainsi que l'analyse et les conclusions.

Une attention particulière doit être accordée aux points suivants:

- la représentation graphique des résultats;
- la présentation graphique de la sélection et la spécification des estimateurs expérimentés choisis;
- la définition de la conception de l'expérience;
- la spécification et le choix des séquences de test;
- les informations générales relatives au système utilisé pour transformer les séquences de test;
- le type de configuration des voies utilisé pour les essais (Recommandation UIT-R BS.775 ou Recommandation UIT-R BS.2051) et description;

  si le système sonore faisant l'objet des essais n'est pas défini dans la Recommandation UIT-R BS.775, la position de chaque haut-parleur du système en question doit être présentée avec un niveau de détail équivalent à celui prévu dans la Recommandation UIT-R BS.775 afin de permettre une reproduction des essais. La position d'écoute de référence par rapport à la position des haut-parleurs associés au système sonore faisant l'objet des essais doit elle aussi être présentée en détail (voir les § 8.5.4 et 8.5.5 de la Recommandation UIT-R BS.1116);
- les détails matériels relatifs à l'environnement et au matériel d'écoute, notamment les dimensions et les caractéristiques acoustiques de la pièce, les types de transducteurs et leur emplacement, les spécifications de l'équipement électrique (voir la Note 1);
- si les exigences relatives aux distances indiquées au § 8.5.1.2 de la Recommandation UIT-R BS.1116 sont respectées. Si ces distances ne sont pas respectées, ce point doit être noté;
- si les exigences relatives aux distances indiquées au § 8.5.1.2 de la Recommandation UIT-R BS.1116 ne sont pas respectées, les méthodes utilisées pour contrôler les réflexions rapides et respecter les exigences données au § 8.3.3.1 de la Recommandation UIT-R BS.1116 devraient être décrites;
- la réponse mesurée du local d'écoute pour tous les haut-parleurs. L'éventuelle application d'une égalisation doit être indiquée, de même que les méthodes utilisées pour ce faire;
- tout écart par rapport aux exigences accoustiques et physiques définies pour le local dans le présent document devrait être indiqué. Ces écarts peuvent concerner les mesures et les réponses accoustiques du local spécifiées au § 8.3 de la Recommandation UIT-R BS.1116, les critères de réponse comportementale de tous les haut-parleurs donnés au § 8.4 de la Recommandation UIT-R BS.1116 et toutes les exigences en matière de distance physiques données au § 8.5 de la Recommandation UIT-R BS.1116;
- la réponse impulsionnelle de chaque haut-parleur, mesurée à toutes les positions d'écoute de l'auditeur, le local étant dans la configuration qui sera utilisée pour l'essai (meubles compris), donnée dans le domaine temporel;
- la conception de l'expérience, la formation, les instructions, les séquences expérimentales, les protocoles d'essai, la production de données;
- le traitement des données, notamment les détails associés aux statistiques inductives descriptives et analytiques;
- l'emploi de repères au cours du test;

– les méthodes de post-sélection qui ont été utilisées lors de l'analyse des résultats, notamment les méthodes d'exclusion des données aberrantes ou l'exclusion des auditeurs non formés;

– l'exécution du test sur la base de la Recommandation UIT-R BS.1534 ou de la Recommandation UIT-R BS.1534-1. Il convient d'indiquer cela clairement dans le document, et de décrire les conditions employées pour les repères;

– la définition appropriée et le code de production dont doit disposer un nouvel utilisateur pour produire un repère employé dans le test, qui n'est pas explicitement décrit dans la présente Recommandation UIT-R BS.1534-2;

– les fondements précis de toutes les conclusions tirées.

NOTE 1 – Puisqu'il semble que les conditions d'écoute, par exemple, la reproduction par haut-parleurs ou celle par casques d'écoute, peut avoir une incidence sur les résultats des évaluations subjectives, les expérimentateurs sont priés de rendre compte explicitement des conditions d'écoute ainsi que du type de l'équipement de reproduction utilisé dans les expériences. Si l'on souhaite effectuer une analyse statistique combinée des différents types de transducteur, il convient de vérifier qu'une telle combinaison des résultats est possible.

10.3 Présentation des résultats

Il faut indiquer pour chaque paramètre de test la médiane et l'intervalle IQR de la distribution statistique des notes d'évaluation.

Les résultats doivent être accompagnés des informations suivantes:

– la description des séquences de test;

– le nombre d'estimateurs;

– une présentation graphique des résultats. Les diagrammes en boîtes représentant les intervalles IQR ainsi que les moyennes et les intervalles de confiance à 95% doivent en faire partie. Il convient de mentionner les différences significatives entre les systèmes testés ainsi que la méthode appliquée pour l'analyse statistique.

Par ailleurs, lorsqu'après la visualisation sous la forme de diagrammes en boîtes les données le permettent, les résultats peuvent aussi être présentés sous des formes appropriées, telles que des moyennes ou des intervalles de confiance.

10.4 Notes absolues

La présentation des notes moyennes absolues pour les systèmes testés, pour la référence cachée et pour les repères donne un bon aperçu des résultats. Il convient toutefois de garder à l'esprit que cela ne contient pas les informations de l'analyse statistique détaillée. En conséquence, les observations ne sont pas indépendantes et l'analyse statistique des notes absolues seules, sans tenir compte de la population sous-jacente des échantillons observés, ne conduirait pas à des informations sensées. En outre, il doit être fait état des méthodes statistiques appliquées, conformément au § 9.

10.5 Niveau de signification et intervalle de confiance

Le rapport de test doit fournir au lecteur des informations relatives à la nature statistique intrinsèque de toutes les données subjectives. Pour faciliter la compréhension du lecteur, il convient d'indiquer les niveaux de signification ainsi que d'autres détails concernant les méthodes statistiques et les résultats. Ces détails pourront inclure les intervalles de confiance ou les barres d'erreur sur les graphiques.

Il n'existe bien sûr pas de niveau de signification «correct». Toutefois, La valeur de 0,05 est généralement choisie. Il est en principe possible d'utiliser un test unilatéral ou bilatéral en fonction de l'hypothèse testée.

Références

[1] Stevens, S. S. (1951). Mathematics, measurement and psychophysics, *in* Stevens, S. S. (éd.), Handbook of experimental psychology, John Wiley & Sons, New York.

[2] UER [2000a] MUSHRA – Method for Subjective Listening Tests of Intermediate Audio Quality. Projet de Recommandation de l'UER, B/AIM 022 (Rev.8)/BMC 607rev, janvier.

[3] UER [2000b] Rapport de l'UER sur les Subjective listening tests of some commercial internet audio codecs. Document BPN 029, juin.

[4] Soulodre, G. A. et Lavoie, M. C. (août 1999). Subjective evaluation of large and small impairments in audio codecs, *in Audio Engineering Society Conference:* 17th International Conference*: High-Quality Audio Coding*, Audio Engineering Society.

[5] Berry, K. J., Johnston, J. E. et Mielke, P. W. (2011). Permutation methods, *Wiley Interdisciplinary Reviews: Computational Statistics*, *3*(6), 527-542.

[6] Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans, *Society of Industrial and Applied Mathematics CBMS-NSF Monographs*, 38.

[7] Cohen, J. (1977). Statistical power analysis for the behavioral sciences (rev. Lawrence Erlbaum Associates, Inc.).

[8] Keppel, G. et Wicken., T. D. (2004). Design and Analysis, *A Researcher's Handbook*, 4ème édition, Pearson Prentice Hall.

[9] Garson, D. G. (2012). Testing statistical assumptions, Blue Book Series, Statistical Associates Publishing.

[10] Ellis, P. D. (2010). The essential guide to effect sizes, *Cambridge: Cambridge University Press*, 3-173.

[11] Howell, D.C. (1997). Statistical methods for psychology, 4ème édition, Duxbury Press.

[12] Kirk, R.E., (1982). Experimental Design: Procedures for the Behavioural Sciences, 2ème édition, Brooks/Cole Publishing Company.

[13] Bech, S. et Zacharov, N. (2007). Perceptual audio evaluation – Theory, method and application, John Wiley & Sons.

[14] Khan, A. et Rayner, G. D. (2003). Robustness to Non-Normality of Common Tests for the Many-Sample Location Problem, *Journal of Applied Mathematics & Decision Sciences*, 7(4), 187-206.

[15] Procédures pratiques d'évaluation subjective de l'UIT-T, Union internationale des télécommunications, 2011.

[16] Faul, F., Erdfelder, E., Buchner, A. et Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses, *Behavior Research Methods, 41,(4),* 1149-1160.

Appendice 1
à l'Annexe 1
(à titre normatif)

Instructions à donner aux estimateurs

Un exemple du type d'instructions qui doivent être données aux estimateurs ou leur être lues afin de les instruire sur la façon d'exécuter le test est donné ci-après.

1 Phase de familiarisation ou de formation

La première étape des tests d'écoute a pour objet la familiarisation avec le protocole d'essai. Cette phase, nommée phase de formation, précède la phase d'évaluation proprement dite.

Le but de la phase de formation est de vous permettre, en tant qu'estimateur, d'atteindre les deux objectifs suivants:

– **Partie A**: vous familiariser avec tous les extraits sonores testés et avec leurs niveaux de qualité;

– **Partie B**: vous apprendre à utiliser l'équipement de test et l'échelle de notation.

Dans la partie A de la phase de formation, vous pourrez écouter tous les extraits sonores qui ont été choisis pour les tests afin de donner une idée des différentes qualités possibles. Les éléments sonores que vous écouterez seront plus ou moins décisifs en fonction du débit binaire et d'autres «critères» employés. La Fig. 3 représente l'interface utilisateur. Vous pourrez, en cliquant sur les différents boutons, écouter différents extraits sonores, y compris les extraits de référence. Vous pourrez ainsi apprendre à apprécier les différents niveaux de qualité des différents éléments de programme. Les extraits sont regroupés en fonction de critères communs. Trois groupes ont ainsi été définis. Chaque groupe comprend quatre signaux transformés.

Dans la Partie B de la phase de formation, vous apprendrez à utiliser les équipements de lecture et de notation mis à votre disposition pour évaluer la qualité des extraits sonores.

Au cours de la phase de formation, vous apprendrez comment, en tant qu'individu, vous pourrez traduire les dégradations audibles en termes de notes. Vous ne devrez jamais, au cours de cette phase, débattre de votre interprétation personnelle des notes avec les autres estimateurs. Vous êtes toutefois encouragé à leur faire connaître les artéfacts.

Lors des tests réels, il ne sera tenu compte d'aucune note attribuée pendant la phase de formation.

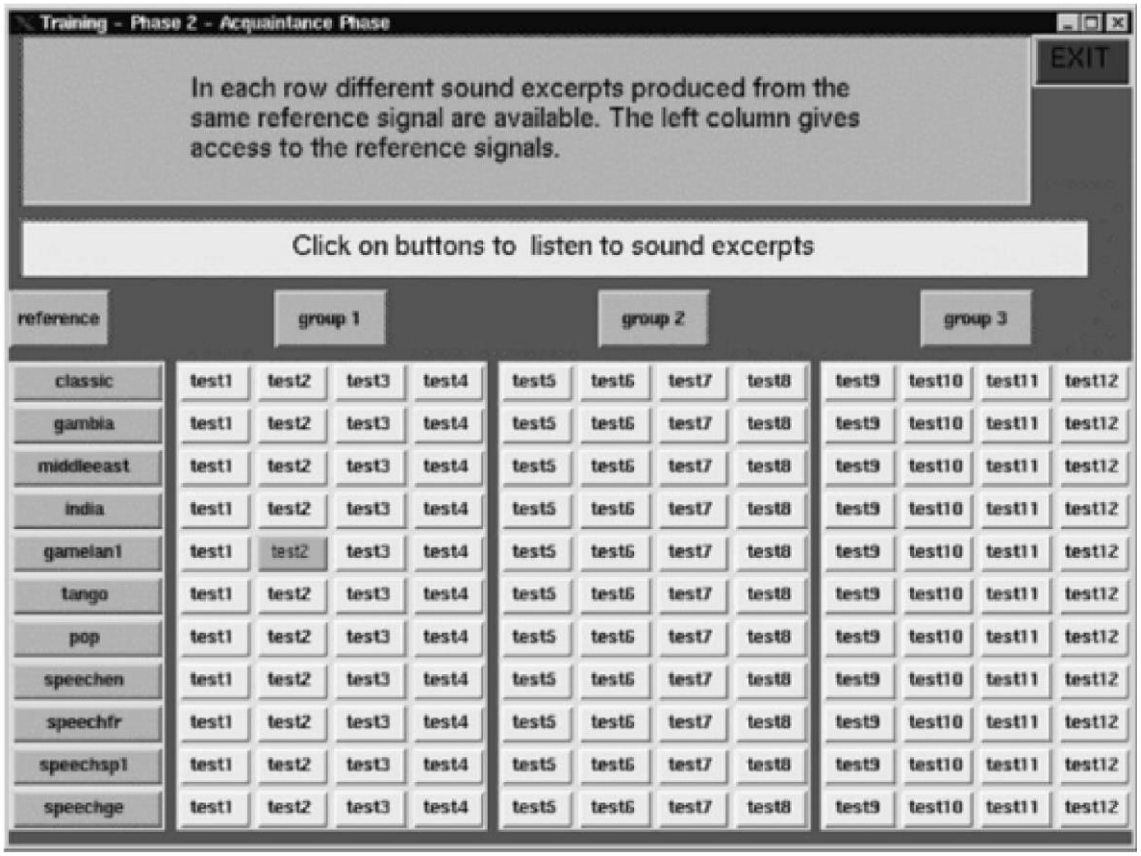
2 Phase de notation en aveugle

La phase de notation en aveugle a pour objet de vous inviter à attribuer vos notes en utilisant l'échelle de qualité. Vos notes doivent traduire votre jugement subjectif concernant le niveau de qualité de chacun des extraits sonores qui vous sont présentés. Chaque essai comportera 9 signaux à noter, dont la durée sera d'environ 10 s. Vous devrez écouter la référence, le repère, ainsi que l'ensemble des conditions de test en cliquant sur les boutons correspondants. Vous pourrez écouter les signaux dans un ordre quelconque et autant de fois que vous le souhaiterez.

Veuillez utiliser le curseur associé à chacun des signaux pour indiquer votre opinion quant à sa qualité. Lorsque vous êtes satisfait des notes que vous avez attribuées à tous les signaux, vous devrez cliquer sur le bouton «register scores» (enregistrer les notes) au bas de l'écran.

FIGURE 3

Exemple d'une interface utilisateur pour la partie A de la phase de formation



BS.1534-03

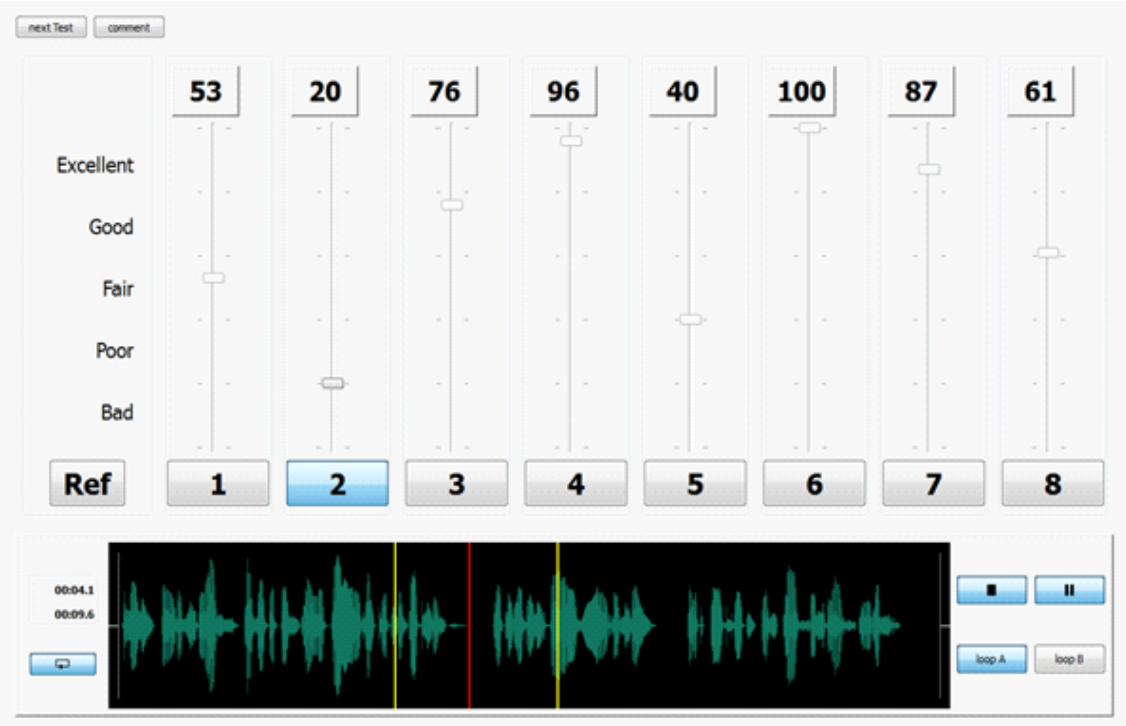
Vous utiliserez pour attribuer vos notes l'échelle de qualité représentée dans la Fig. 1.

L'échelle de notation est continue, allant de la catégorie «excellent» à la catégorie «mauvais». La note 0 correspond à la note minimale de la catégorie «mauvais», tandis que la note 100 correspond à la note maximale de la catégorie «excellent».

Lors de l'évaluation des extraits sonores, veuillez noter que vous ne devez pas nécessairement attribuer une note dans la catégorie «mauvais» à l'extrait sonore du test dont la qualité est la plus faible. En revanche, vous devez attribuer la note 100 à un ou à plusieurs extraits, parce que le signal de référence non transformé fait partie des extraits à évaluer.

FIGURE 4

Exemple d'une interface utilisateur employée pendant la phase de notation en aveugle



BS.1534-04

Appendice 2
à l'Annexe 1
(à titre informatif)

Notes d'orientation en matière de conception d'une interface utilisateur

Les suggestions suivantes sont faites pour ceux qui envisageraient:

a) de mettre au point des systèmes permettant d'exécuter des tests subjectifs selon la méthode MUSHRA;

b) de réaliser de tels tests.

Elles visent à accroître la fiabilité des résultats de test et à faciliter l'analyse de toutes les irrégularités susceptibles d'être observées au cours du traitement des notes d'évaluation.

La conception de l'interface utilisateur doit être telle que le risque qu'un sujet attribue une note qui ne corresponde pas à sa véritable intention soit minimisé. A cette fin, des mesures doivent être prises pour garantir que l'interface utilisateur indique clairement quelle version transformée d'un élément de test est écoutée par le sujet à un moment donné. Cela peut être facilité par un choix soigneux des couleurs et de la luminosité des indicateurs sur l'écran (les boutons sur lesquels on peut cliquer, par exemple) de manière à éviter les difficultés pouvant se présenter lorsqu'un sujet est non sensible à certaines couleurs.

Il faut aussi faire en sorte que le sujet soit en mesure de modifier uniquement la note qu'il a attribuée à l'élément en cours d'écoute. On a observé que certains estimateurs écoutent successivement deux versions transformées d'un même élément afin d'attribuer une note à la première version qu'ils ont entendue, et non à la seconde. Dans ce cas, il est possible qu'une erreur soit commise (en particulier lorsqu'un grand nombre de commandes sont affichées sur l'écran), la note pouvant être attribuée à un signal autre que le signal voulu. Pour que cela se produise moins, il est suggéré que la seule commande activée à un instant donné soit celle qui est associée au signal en cours d'écoute. Les commandes, qui permettent d'attribuer des notes à des signaux ne faisant pas l'objet d'une écoute, doivent être désactivées.

Appendice 3
à l'Annexe 1
(à titre normatif)

Description d'une comparaison statistique non paramétrique de deux échantillons au moyen des techniques de rééchantillonnage et des méthodes de simulation de Monte-Carlo

Les tests non paramétriques de simulation de distribution aléatoire peuvent être employés avec les techniques de rééchantillonnage courantes, telles que les procédures de «bootstrap», pour déterminer la signification de presque tous les résultats statistiques. Par exemple, la signification d'une différence en matière de réponse entre les médianes observées pour deux signaux de test (la taille des échantillons étant $N1$ et $N2$) peut se calculer comme suit. La différence effective entre les médianes de chacun des échantillons doit être notée et désignée par $Diff_{ACT\_1}$. Toutes les données provenant de ces échantillons doivent ensuite être agrégées en une seule suite ou un seul vecteur. Une procédure de «bootstrap» doit ensuite être employée de manière que, à chaque itération, l'ensemble agrégé soit permuté, les échantillons de taille $N1$ et $N2$ étant tirés sans remplacement. La différence entre les médianes des deux échantillons tirés de manière aléatoire doit être enregistrée en tant que $Diff_{EST\_1}$. Cette procédure peut ensuite être reproduite 10 000 fois. Le rapport entre le nombre de fois où $Diff_{EST\_N}$ dépasse $Diff_{ACT\_N}$, divisé par 10 000, donne une valeur p correspondante. Si le nombre total de fois où $Diff_{EST\_N}$ dépasse $Diff_{ACT\_N}$ est inférieur à 500 (500/10 000 = 0,05), alors la différence entre les deux médianes peut être considérée comme significative avec un niveau de 0,05, $p < 0,05$.

Appendice 4
à l'Annexe 1
(à titre informatif)

Notes d'orientation en matière d'analyse statistique paramétrique

1 Introduction

L'analyse statistique paramétrique des résultats obtenus lors des tests MUSHRA est décrite au § 9. Toutefois, lorsque de nombreuses conditions doivent être comparées entre-elles, il est préférable, au lieu d'effectuer plusieurs comparaisons par paires, d'employer un test omnibus tel que l'analyse ANOVA. Le présent Appendice décrit comment cela peut se faire. Il contient aussi les conditions préalables à l'analyse et signale des alternatives lorsque celles-ci ne sont pas satisfaites.

Le test MUSHRA fait appel à un *plan à mesures répétées* ou *intra-sujets* (une excellente introduction à ces concepts est donnée dans l'ouvrage de Maxwell et Delaney (2004)), dans lesquelles deux facteurs intra-sujets (condition et séquence audio) sont entièrement croisés et au moins une note est obtenue pour chaque combinaison (auditeur, séquence audio et condition). Il peut aussi y avoir des cas où les mêmes combinaisons (séquence audio et condition) sont soumises à deux ou plus de deux groupes différents d'estimateurs, par exemple, dans différents laboratoires. Dans ces cas, il convient de tenir compte dans l'analyse du facteur supplémentaire inter-sujets *groupe*.

Les statistiques inductives sont nécessaires à la généralisation des résultats obtenus pour un échantillon comparativement petit de la population des auditeurs. Par exemple, si lors du test d'écoute la notation indique la présence d'une différence entre la qualité audio perçue avec un nouveau codeur et celle qui est perçue avec un codeur existant, il est important de répondre à la question de savoir si l'on peut aussi s'attendre à cette différence dans le cas où la qualité audio des deux systèmes aurait été notée par un groupe entièrement différent d'auditeurs. En ce qui concerne la conception proprement dite des tests d'écoute MUSHRA, il y a au moins trois questions auxquelles l'on peut vouloir répondre (ou, en termes statistiques, des hypothèses que l'on souhaite tester), et les statistiques inductives décrites ici apportent des réponses valables. Premièrement, la question primordiale sera généralement celle de savoir si la qualité audio perçue diffère d'un système testé à un autre (par exemple, la référence et trois codeurs différents). Deuxièmement, si au cours du test d'écoute les systèmes audio sont évalués avec des séquences de test différentes, les notes de la qualité audio dépendent-elles de la séquence audio? Troisièmement, l'effet du système audio sur la qualité audio perçue diffère-t-il d'une séquence de test à l'autre? Pour répondre comme il convient à ces questions, il faut d'abord tester la signification de l'effet majeur de la condition (système audio), de l'effet majeur de la séquence audio et de l'interaction condition × séquence audio, en effectuant une analyse de variance (ANOVA). Il y a interaction lorsque les différences entre les qualités perçues des systèmes audio dépendent de la séquence audio. Il convient de noter qu'en raison des interactions possibles il n'est pas conseillé d'agréger les notes pour un système audio, obtenues avec différentes séquences audio, même si l'on ne s'intéresse pas particulièrement à l'effet de la séquence audio ou des interactions. Des hypothèses plus spécifiques, concernant par exemple la différence perçue entre deux systèmes audio, peuvent ensuite être testées au moyen de comparaisons supplémentaires.

Lorsque plus de deux conditions expérimentales doivent être comparées, par exemple quatre codeurs différents, il n'est pas indiqué que la statistique inductive soit fondée sur de multiples comparaisons par paires. Par exemple, si $K = 5$ systèmes audio sont utilisés dans le test (4 codeurs et une référence), le nombre de paires de conditions est égal à $\binom{K}{2} = K(K-1)/2 = 10$. L'analyse visant à détecter les différences pour chacune de ces 10 paires au moyen de 10 tests *t* pour échantillons appariés à un niveau α de 0,05 conduira à une inflation du taux d'erreurs de type I par famille. Pour chacun des tests *t*, la probabilité d'un rejet erroné de l'hypothèse nulle selon laquelle il n'y a pas de différence entre les qualités audio perçues des deux codeurs est α.

Parmi ces C tests, la probabilité de commettre au moins une erreur de type I est $1 - (1 - \alpha)^C$, qui pour $C = 10$, comme dans notre exemple, est égale à 0,40 et est donc bien supérieure au niveau α souhaité de 0,05. Le taux d'erreurs par famille peut être réduit en apportant des corrections appropriées aux tests multiples, telles que la correction de Bonferroni ou la procédure de Hochberg (1988) décrites plus loin. Mais les tests *t* par paires avec correction dissimulent des informations pertinentes, en partie parce que plusieurs tests *t* sur toutes les paires de moyennes emploient des informations redondantes (chaque moyenne apparaissant dans plusieurs tests). La méthode des tests par paires est généralement moins puissante (c'est-à-dire moins sensible en ce qui concerne la détection de différences entre les conditions) que l'emploi d'un test omnibus approprié, qui, dans le cas du test MUSHRA, consiste à effectuer une analyse de variance à mesures répétées (rmANOVA). L'analyse des données dans le cas d'un test MUSHRA où n'interviennent pas de facteurs inter-sujets est décrite pas à pas ci-après. En d'autres termes, on suppose qu'un seul groupe d'estimateurs a été testé et que toutes les combinaisons (condition et séquence audio) ont été soumises au moins une fois à chacun des estimateurs. L'extension à un plan faisant intervenir plusieurs groupes (par exemple, lorsque le test a été exécuté dans deux laboratoires) sera décrite plus loin.

2 Test visant à détecter la normalité

Il est prudent de tenir compte des effets sur la validité du test statistique d'une déviation possible par rapport à la normalité de la valeur des réponses. S'agissant d'un plan intra-sujets, dans laquelle chaque estimateur testé est confronté à une seule condition expérimentale, les analyses ANOVA effectuées dans le cadre du modèle linéaire général sont étonnamment robustes par rapport à la non-normalité de la valeur des réponses (voir par exemple les références [11]; [13]; [25] et [35]).

Pour un plan à mesures répétées, comme dans le test MUSHRA, il faut noter en premier lieu qu'il existe une autre façon de tester l'hypothèse nulle selon laquelle dans la population la qualité audio perçue est la même pour toutes les conditions. Cela équivaut à calculer $K - 1$ contrastes orthogonaux, par exemple en définissant des variables différence entre les K conditions, puis en testant l'hypothèse selon laquelle la moyenne sur la population de ces variables différence est égale à 0. Par exemple, si les tests comportent la référence et deux codeurs, alors deux variables différence D_1 et D_2 peuvent être obtenues en calculant pour chaque sujet la différence entre la note de la référence et la note du codeur A (D_1), et la différence entre la note du codeur A et la note du codeur B (D_2). Les méthodes ANOVA à mesures répétées supposent toutes que ces variables différence sont distribuées selon une loi normale multidimensionnelle. Malheureusement, contrairement à ce qui vaut pour le plan inter-sujets, la non-normalité peut conduire à des taux d'erreurs de type I trop prudents ou trop présomptueux ([5]; [22]; [30] et [39]). Cela veut dire que pour un niveau α donné (par exemple $\alpha = 0{,}05$), bien que l'hypothèse nulle selon laquelle des moyennes identiques pour toutes les conditions soit vraie, la proportion des cas dans lesquels l'analyse ANOVA produit une valeur *p* significative ($p < \alpha$) sera inférieure ou supérieure à la valeur nominale α. A nouveau, contrairement à ce qui vaut pour le plan inter-sujets, la simple augmentation de la taille des échantillons ne résout

pas ce problème [30]. Il est de plus en plus évident qu'en termes d'aplatissement les écarts en matière de symétrie ont un effet plus important que les déviations par rapport à la distribution normale ([4] et [18]). Le degré de déviation par rapport à la symétrie peut s'exprimer en termes de l'*asymétrie* de la distribution, qui est égale au moment normalisé d'ordre trois [8]. Pour une distribution symétrique telle que la distribution normale, l'asymétrie est nulle. L'*aplatissement*, à savoir le moment normalisé d'ordre quatre par rapport à la moyenne, décrit la pondération entre le pic et la queue (voir la référence [9] pour des illustrations). De précédentes études de simulation indiquent que pour de petites déviations par rapport à la symétrie, l'analyse rmANOVA permettra de réduire encore le taux d'erreurs de type I. Toutefois, l'état actuel de la recherche ne permet pas de formuler des règles précises concernant le degré acceptable de déviation par rapport à la normalité. Il est donc recommandé d'exécuter des tests visant à détecter la normalité multidimensionnelle de la distribution et de consigner les estimations empiriques de l'asymétrie et de l'aplatissement.

Il est important de noter que, dans le modèle linéaire général sous-jacent à l'analyse rmANOVA, les réponses brutes (à savoir les notes du test MUSHRA) ne sont pas supposées être sont normalement distribuées. Au contraire, on suppose dans le modèle que les *erreurs* sont normalement distribuées. Pour cette raison, les tests de la normalité ou les valeurs de l'asymétrie et de l'aplatissement doivent être calculées pour les *résidus* du modèle, plutôt que pour les données brutes. Heureusement, la plupart des logiciels statistiques sont capables d'enregistrer les résidus pour chacune des conditions expérimentales analysées, qui dans le cas présent sont les différentes combinaisons (système audio et séquence audio). Un vecteur de résidus est ainsi obtenu pour chacune des conditions expérimentales. Les composantes de chaque vecteur correspondent à un estimateur.

Plusieurs tests de la normalité multidimensionnelle sont disponibles, par exemple, le test multidimensionnel de Shapiro-Wilk proposé par Royston [34], les tests fondés sur l'asymétrie et l'aplatissement multidimensionnels [10] et d'autres encore [14]. Des macros d'exécution de ces tests sont à disposition dans les logiciels SPSS (http://www.columbia.edu/~ld208/normtest.sps) et SAS (http://support.sas.com/kb/24/983.html), ainsi que très probablement dans d'autres logiciels. Des estimations unidimensionnelles de l'asymétrie et de l'aplatissement, qui peuvent être calculées séparément pour les résidus de chacune des combinaisons (système audio et séquence audio), sont effectuées par tous les principaux logiciels de statistique. La macro SPSS de DeCarlo [9] (http://www.columbia.edu/~ld208/normtest.sps) permet également de calculer l'asymétrie et l'aplatissement multidimensionnels [26]. Les estimations de l'asymétrie et de l'aplatissement unidimensionnels ou multidimensionnels doivent être consignées, ainsi que le résultat du test de la normalité multidimensionnelle.

Si le test de la normalité multidimensionnelle n'est pas significatif, ou si tous les tests multidimensionnels ou unidimensionnels n'indiquent pas de déviation significative de l'asymétrie et de l'aplatissement par rapport aux valeurs escomptées pour une distribution normale, alors les hypothèses de l'analyse rmANOVA sont satisfaites.

Si toutefois l'un des tests indique une déviation significative par rapport à la normalité, ou si l'asymétrie pour une condition expérimentale quelconque dépasse la valeur de 0,5 (en tant que règle empirique provisoire), la question de savoir quelles en sont les conséquences se pose alors. Il y a deux problèmes d'ordre général, tous deux associés à l'absence susmentionnée de règles relatives à une déviation acceptable par rapport à la normalité dans les analyses rmANOVA. Premièrement, les tests de la normalité multidimensionnelle sont plutôt sensibles et détecteront souvent de très petites déviations par rapport à la normalité. Ils détecteront aussi, non seulement une asymétrie dans la distribution des résidus, mais aussi l'aplatissement ou d'autres aspects de la distribution, alors que très probablement seule l'asymétrie conduit à des taux d'erreurs de type I non robustes dans les analyses rmANOVA. Deuxièmement, si les valeurs de l'asymétrie et de l'aplatissement multidimensionnels sont estimées à partir des données [26], cette information ne permet pas de décider si l'analyse

rmANOVA peut être effectuée, à nouveau en raison de l'absence de règles relatives à une déviation acceptable par rapport à la normalité. Cela souligne la nécessité de consigner les valeurs de l'asymétrie et de l'aplatissement ainsi que les résultats de test. Dès que des règles valables concernant une déviation acceptable par rapport à la normalité seront disponibles, les résultats des tests rmANOVA pourront alors être réévalués sur la base de meilleures informations. Si la déviation par rapport à la normalité semble forte, indiquée par exemple par des estimations de l'asymétrie supérieures à 1,0 [29], des alternatives non paramétriques à l'analyse rmANOVA pourraient être envisagées, comme, par exemple, des tests employant les techniques de rééchantillonnage ou le test de Friedman. Toutefois, il reste à déterminer dans quelles situations les techniques de rééchantillonnage permettent de résoudre le problème de la non-normalité [38]. Dans le test de Friedman, la normalité multidimensionnelle n'est pas supposée, mais on suppose que les variances sont identiques pour toutes les conditions expérimentales [36], ce qui ne sera souvent pas le cas pour les données expérimentales. De plus, le test de Friedman est un test unidimensionnel. En raison de cela, même si l'hypothèse de variances égales est satisfaite, ce test peut être employé pour détecter l'effet moyen du système audio sur les séquences audio, mais il ne peut être utilisé pour analyser l'interaction système audio × séquence audio.

3 Choix de l'analyse de variance à mesures répétées

S'agissant de plans à mesures répétées, de nombreuses approches peuvent être employées pour tester les effets des facteurs intra-sujets et des facteurs inter-sujets [21]. Puisque nous étudions ici le cas d'un plan qui ne contient pas de facteur inter-sujets (groupe), et que nous supposons qu'aucune donnée ne manque (c'est-à-dire qu'une note est disponible pour chacune des combinaisons (auditeur, séquence audio et condition)), nous pouvons recommander deux approches. Ces approches fournissent toutes les deux des tests des hypothèses dans le cas où la distribution des données est une distribution normale multidimensionnelle, mais, en fonction entre autres de la taille de l'échantillon, elles peuvent différer pour ce qui est de leur puissance statistique (par exemple, en ce qui concerne la sensibilité nécessaire pour détecter un écart par rapport à l'hypothèse nulle).

Les deux variantes d'analyse sont a) l'*approche unidimensionnelle avec la correction de Huynh-Feldt pour les degrés de liberté*, et b) l'*approche multidimensionnelle*. Des descriptions détaillées de ces approches sont données ailleurs ([21] et [28]). Les deux variantes sont incorporées dans les principaux logiciels statistiques (par exemple, les logiciels R, SAS, SPSS et Statistica).

En raison de la structure à mesures répétées des données, les notes obtenues pour les différentes combinaisons (condition et séquence audio) sont corrélées. Par exemple, si un auditeur attribue une note inhabituellement élevée à un repère de faible qualité, alors les notes qu'il attribuera aux codeurs auront aussi tendance à être plus élevées que les notes des autres estimateurs. L'approche unidimensionnelle suppose que la structure de variance-covariance des données est sphérique, ce qui revient à dire que les variables différence décrites ci-après ont toutes la même variance ([16] et [33]). Toutefois, cette hypothèse est violée pour quasiment tous les ensembles de données empiriques [21]. Pour résoudre ce problème, un facteur de correction est appliqué aux degrés de liberté lors du calcul de la valeur *p* selon la distribution *F*. A ces fins, l'écart par rapport à la sphéricité est évalué à partir des données. Le facteur de correction de Huynh-Feldt, nommé $\tilde{\varepsilon}$, est recommandé [17] parce qu'un autre facteur de correction, le facteur de Greenhouse-Geisser [12] a tendance à conduire à des tests prudents (par exemple, les références [17] et [30]). Lorsque la distribution des données est normale, l'approche unidimensionnelle avec la correction de Huynh-Feldt produit des taux d'erreurs de type I acceptables, même pour des échantillons extrêmement petits ($N = 3$). Le facteur de correction $\tilde{\varepsilon}$ et les valeurs *p* corrigées sont incorporés dans tous les logiciels statistiques.

L'*approche multidimensionnelle* emploie une formulation différente mais équivalente de l'hypothèse nulle. Par exemple, considérons l'hypothèse nulle selon laquelle dans la population la qualité audio perçue est identique pour toutes les conditions. Cela équivaut à calculer $K - 1$ contrastes orthogonaux, par exemple, en définissant des variables différence entre les K conditions, puis en testant l'hypothèse selon laquelle le vecteur μ des moyennes de la population de tous les $K - 1$ contrastes est égal au vecteur nul ($\mu = 0$). Par exemple, si la référence et deux codeurs sont soumis, alors deux variables différence D_1 et D_2 peuvent être obtenues en calculant pour chaque estimateur la différence entre la note de la référence et la note du codeur A (D_1), et la différence entre la note du codeur A et la note du codeur B (D_2). L'analyse rmANOVA employant l'approche multidimensionnelle est fondée sur les variables différence et fait appel à un test multidimensionnel de l'hypothèse $\mu = 0$. Dans cette approche, aucune hypothèse concernant la matrice des variances-covariances n'est nécessaire. Lorsque la distribution des données est une distribution normale multidimensionnelle, ce test est exact, mais il nécessite un nombre d'estimateurs au moins aussi grand que le nombre de niveaux des facteurs. En raison de cela, il ne peut être utilisé si, par exemple, 9 conditions (8 codeurs et une référence) ont été soumises à seulement 8 estimateurs.

La puissance relative des deux approches dépend, parmi de nombreux autres facteurs, de la taille des échantillons et du nombre de niveaux du facteur intra-sujets. Selon Algina et Keselman (1997), une simple règle de sélection consisterait à utiliser l'approche unidimensionnelle avec la correction de Huynh-Feldt si $\tilde{\varepsilon} > 0.85$ et $N < K + 30$, où N est le nombre d'estimateurs et K est le nombre maximal de niveaux du facteur intra-sujets. Dans les autres cas, il est préférable d'employer l'approche multidimensionnelle. Il convient de noter que, si l'expérience est conduite dans différents laboratoires, le nombre N est le nombre total d'estimateurs ayant participé à l'étude (par exemple, 10 estimateurs dans le laboratoire A et 10 estimateurs dans le laboratoire B correspondent à $N = 20$).

4 Exécution de l'analyse de variance à mesures répétées choisie et tests a posteriori facultatifs

Au cours de cette étape, des tests omnibus des effets des conditions, des séquences audio et de leur interaction sont exécutés en employant la version de l'analyse rmANOVA choisie. Pour procéder à cette analyse rmANOVA, la plupart des logiciels tels que les logiciels SAS, SPSS et Statistica demandent que les données soient disponibles sous la forme d'«une ligne par estimateur». Donc, le tableau de données ne doit contenir qu'une seule ligne par estimateur et les notes de toutes les combinaisons (condition et séquence audio) sont présentées sous la forme de colonnes (les «variables»).

L'analyse rmANOVA à deux facteurs donne des informations sur trois effets.

1) Effet majeur des conditions

Dans la plupart des cas, il s'agit du test présentant le plus d'intérêt. Si l'analyse ANOVA indique un effet significatif de la condition, alors l'hypothèse nulle selon laquelle dans la population la qualité audio perçue est identique pour toutes les conditions (référence, codeurs 1 à *k*) peut être rejetée. En d'autres termes, le test indique que dans la population il y a des différences entre la qualité audio perçue des systèmes audio. Il n'est pas possible d'utiliser, en tant que mesure de la signification de l'effet, la valeur *d* de Cohen [6] ou une valeur analogue, parce que *d* n'est pas défini pour une comparaison de plus de deux moyennes. Dans un contexte ANOVA, il est courant de mesurer la force de l'association. Ces mesures donnent des informations sur la fraction de la variance des données, qui est imputable à l'effet considéré. C'est la même logique que celle qui est sous-jacente au coefficient de détermination R^2. La plupart des logiciels statistiques sont en mesure de calculer le coefficient η^2 partiel, qui s'obtient comme le rapport entre la variance due à l'effet et la somme de la variance due à l'effet et de la variance due à l'erreur (résiduelle). Une analyse des autres mesures de la force de l'association est donnée dans l'article d'Olejnik et Algina [31].

Après avoir obtenu un résultat de test significatif pour un effet majeur, il sera souvent intéressant de connaître les origines de cet effet. Cela peut se faire en calculant des contrastes spécifiques. Par exemple, il pourrait être intéressant de savoir si la qualité sonore d'un nouveau codeur diffère de celle des trois codeurs existants. Pour répondre à cette question, il faut d'abord calculer la note moyenne attribuée aux trois codeurs existants par chacun des estimateurs, en prenant la moyenne sur les séquences audio. Il y aura donc, pour chaque estimateur a) une note pour le nouveau codeur et b) une note moyenne pour les trois autres codeurs. Ces deux valeurs sont ensuite comparées au moyen d'un test *t* pour échantillons appariés. Il convient de noter que, comme les données proviennent d'un plan à mesures répétées, il est important de ne pas regrouper les variances [27]. Il convient aussi d'observer qu'au lieu d'effectuer l'analyse ANOVA ce contraste aurait pu être testé comme un contraste planifié. Il est généralement recommandé d'employer des tests bilatéraux de signification. Toutefois, s'il y avait par exemple une hypothèse *a priori* selon laquelle le nouveau codeur devrait recevoir de meilleures notes que les codeurs existants, on pourrait admettre l'emploi d'une région de rejet unilatérale.

D'autres contrastes spécifiques peuvent être calculés sur la base de la même logique. Une façon de tester les contrastes consiste à calculer une combinaison linéaire des notes obtenues dans les différentes conditions expérimentales, puis à employer un test *t* à un échantillon pour décider si ce contraste diffère significativement de 0. Pour chaque estimateur *i*, on calcule un contraste:

$$\Psi_i = \sum_{j=1}^{a} c_j Y_{ij}, \quad \sum_{j=1}^{a} c_j = 0,$$

où Y_{ij} est la note attribuée par l'estimateur *i* dans la condition *j* (en prenant la moyenne sur les séquences audio), *a* est le nombre de conditions dont il est tenu compte dans ce contraste et les c_j sont les coefficients. Dans l'exemple ci-dessus, si le nouveau codeur correspond à $j = 1$ et que les trois autres codeurs correspondent à $j = 2 \ldots 4$, le choix de $c_1 = -1$ et de $c_2 = c_3 = c_4 = 1/3$ permettra de tester l'hypothèse selon laquelle la qualité audio du nouveau codeur diffère de celle des trois autres codeurs.

Si plus d'un contraste *a posteriori* est calculé, alors, comme indiqué ci-dessus, cela pose des problèmes pour les tests multiples. Pour résoudre cela, il est recommandé d'appliquer, comme décrit par Hochberg [15], la procédure montante d'acceptation séquentielle de Bonferroni. Cette procédure contrôle le taux d'erreurs de type I par famille, tout en étant plus puissante que de nombreuses autres procédures [20]. Dans la procédure de Hochberg, on calcule d'abord les *m* contrastes présentant un intérêt et on les ordonne par rapport à la valeur *p*. On commence ensuite à examiner la plus grande valeur *p*. Si cette valeur *p* est plus petite que α, alors toutes les hypothèses sont rejetées (ce qui veut dire que tous les contrastes sont significatifs). Si ce n'est pas le cas, le test *t* avec la valeur *p* la plus grande n'est pas significatif et l'on poursuit en comparant la valeur *p* plus petite suivante avec α/2. Si

cette valeur p est plus petite que $\alpha/2$, alors ce test et tous les tests avec une valeur p plus petite sont significatifs. Si ce n'est pas le cas, le test t avec la deuxième plus grande valeur p n'est pas significatif et l'on poursuit en comparant la valeur p plus petite suivante avec $\alpha/3$. En termes plus formels, si p_i avec $i = m, m - 1, \ldots, 1$ sont les valeurs p en ordre décroissant, alors pour tout i, si $p_i < \alpha/(m - i + 1)$, tous les tests $i' \leq i$ sont significatifs.

En principe, il est aussi possible d'effectuer des comparaisons par paires *a posteriori* entre les notes pour toutes les conditions. Pour un plan à mesures répétées, cela nécessiterait d'exécuter des tests t pour échantillons appariés pour toutes les paires de conditions. Cette approche n'est toutefois pas recommandée. Considérons une expérience avec 7 codeurs et une référence. Pour cet ensemble de 8 conditions, $8 \cdot 7/2 = 28$ tests par paires peuvent être exécutés, et il ne sera pas facile d'extraire des informations utiles d'un aussi grand nombre de tests. Si toutes les différences pour les paires sont testées, alors, en raison du nombre élevé de tests, l'application de la procédure de Hochberg [15] pour corriger les tests multiples est bien sûr spécialement importante. Il convient de noter que, s'il est évident qu'il y a une déviation par rapport à la normalité des différences sur lesquelles le test t pour échantillons appariés est fondé, il existe un autre test ne faisant pas l'hypothèse de la normalité, à savoir le test des signes.

Il convient aussi de noter qu'à la suite d'un effet majeur significatif, il se pourrait qu'aucun des contrastes ou qu'aucune des différences pour les paires *a posteriori* ne soit significatif [28], en raison de la différence des informations statistiques employées par l'analyse rmANOVA et par les tests *a posteriori*. Et fait important, l'analyse rmANOVA est le test qui convient le mieux. Donc, un effet majeur indiqué par l'analyse ANOVA est acceptable même si aucun des tests *a posteriori* ne s'avère être significatif. Si à la suite d'un test omnibus significatif (ANOVA), aucun contraste *a posteriori* n'est significatif, alors on peut conclure que les systèmes audio diffèrent pour ce qui est de la qualité sonore perçue. Les différences entre les systèmes audio peuvent aussi être comparées les unes avec les autres. Par exemple, dans le cas des paires de systèmes audio pour lesquelles la différence entre les notes concernant la qualité sonore est la plus grande, il se peut que ces différences pour les paires deviennent significatives pour une taille d'échantillon plus grande. Toutefois, il doit être conclu que dans cette étude aucune des différences pour les paires n'a été significative.

Si l'analyse rmANOVA ne conduit *pas* à un effet majeur significatif de la condition, cela indique que les différences entre les systèmes soumis au test sont faibles. Toutefois, en raison de la taille finie des échantillons, on ne peut conclure que dans la population il n'y a *pas* de différence entre les conditions pour ce qui est de la qualité audio [3]. Les différences dans la population pourraient être nulles, ou, compte tenu de la taille des échantillons, les effets pourraient avoir été trop petits pour être détectés. Si une analyse de puissance *a priori* avait été conduite, c'est-à-dire si la taille de l'échantillon avait été choisie suffisamment grande pour détecter un effet de dimension donnée avec une probabilité donnée, il peut être conclu que les données témoignent en faveur de l'absence d'un effet de la dimension donnée *a priori*.

Ce résultat pourrait être considéré comme une définition de la transparence des codeurs. Si aucune analyse de puissance *a priori* n'a été effectuée, il faudrait se garder de conclure que les codeurs sont transparents, pour les raisons indiquées ci-dessus. Une solution approximative *a posteriori* usuelle consiste à comparer la valeur p à [0,2] plutôt qu'à 0,05. Si le test reste non significatif, c'est une indication un peu plus forte de l'absence de différences dans la qualité audio perçue des conditions.

2) Effet majeur des séquences audio

En suivant les mêmes étapes et la même logique que ci-dessus, le test de l'effet majeur de la séquence audio fournit des informations sur les variations systématiques des notes en fonction des séquences de test. Pour la plupart des scénarios de test MUSHRA, cet effet ne devrait pas présenter beaucoup d'intérêt, parce qu'il n'a pas de rapport direct avec les différences entre les systèmes audio.

3) Interaction entre les conditions et les séquences audio

Si l'analyse rmANOVA indique une interaction significative entre les conditions et les séquences audio, alors l'effet du système audio sur la qualité audio perçue diffère selon les séquences de test. Par exemple, la référence et un codeur pourraient être notés de la même façon pour un chant pop fortement comprimé, où les artéfacts de codage sont masqués par la distorsion inhérente de la séquence. D'autre part, la note de la qualité sonore pour le codeur pourrait être inférieure à celle qui est attribuée à la référence, s'agissant de l'enregistrement à plage dynamique étendue d'un piano de concert. Cette interaction présente en général de l'intérêt pour un test MUSHRA, parce qu'elle indique que la différence entre les systèmes audio dépend de la séquence de test.

A la suite d'un test omnibus significatif de l'effet de l'interaction, la nature de l'interaction peut encore être examinée au moyen de tests *a posteriori*. Une approche courante consiste à tester les *effets majeurs simples*. Ceux-ci peuvent, par exemple, être calculés en effectuant plusieurs analyses rmANOVA distinctes à un facteur avec la condition de facteur intra-sujets, une pour chacune des séquences audio. Ces analyses indiqueront les séquences audio pour lesquelles l'effet de la condition est majeur. A nouveau, la procédure de Hochberg devrait être appliquée pour corriger les tests multiples.

Comme ci-dessus, toutes les différences pour les paires de combinaisons (condition et séquence audio) pourraient en principe être testées au moyen de tests *t* pour échantillons appariés distincts et de la procédure de Hochberg. Le nombre de comparaisons par paires sera toutefois encore plus grand que dans le cas des effets majeurs. Si par exemple 8 systèmes audio sont combinés avec 4 séquences de test, il y a 24 combinaisons (système audio et séquence de test), correspondant à $24 \cdot 23/2 = 276$ tests par paires. A l'évidence, cette approche ne peut être recommandée.

5 Extension à des plans contenant une variable inter-sujets (groupe)

Jusqu'à présent, nous avons considéré un plan sans facteurs inter-sujets. Quelles analyses conviendrait-il d'effectuer lorsque le test a été exécuté sur différents groupes d'estimateurs, par exemple, dans deux laboratoires ou avec des musiciens et des non-musiciens?

Si des facteurs inter-sujets sont présents, il est extrêmement important de savoir si le nombre d'estimateurs est le même dans tous les groupes (plan équilibré) ou si ce nombre diffère d'un groupe à l'autre (plan déséquilibré).

Plan équilibré. Si le nombre d'estimateurs est identique à tous les niveaux du facteur inter-sujets, ou si les tailles des groupes ne diffèrent pas de plus de 10%, alors, à nouveau, l'analyse rmANOVA peut être effectuée soit au moyen de l'approche unidimensionnelle avec la correction de Huynh-Feldt pour les degrés de liberté, soit au moyen de l'approche multidimensionnelle [21]. Le plan contiendra maintenant des facteurs intra-sujets condition et séquence audio et au moins un facteur inter-sujets (par exemple, les laboratoires). En raison de cela, l'analyse rmANOVA fournira un test supplémentaire du ou des effets inter-sujets ainsi que des interactions entre tous les effets intra- et inter-sujets.

Par exemple, il se pourrait que l'interaction condition × laboratoire soit significative, ce qui voudrait dire que la qualité audio des systèmes audio perçue dans le laboratoire A diffère de celle qui est perçue dans le laboratoire B. Il convient de noter que nous supposons ici que les deux groupes ont été soumis exactement aux mêmes combinaisons (condition et séquence audio). Si par exemple des séquences audio différentes avaient été présentées dans les deux laboratoires, les méthodes proposées ici ne pourraient pas être utilisées. Au lieu de cela, il faudrait faire appel aux modèles nommés modèles à effets aléatoires [28], qui sortent du cadre du présent Appendice.

Plan déséquilibré. Si les tailles des groupes diffèrent de plus de 10%, alors, malheureusement, ni l'approche unidimensionnelle, ni l'approche multidimensionnelle ne conduisent plus à des résultats de test valables [21]. En conséquence, il est fortement recommandé de prévoir des groupes de tailles égales et donc d'éviter ce problème. Si les tailles des groupes ne sont pas les mêmes, deux procédures d'analyse peuvent être recommandées. La première approche est le test par approximation générale améliorée (IGA) [1], et la deuxième approche est une variante particulière d'une analyse au moyen d'un modèle mixte fondé sur la probabilité maximale [23]. Le test IGA est disponible sous la forme d'une macro dans le logiciel SAS. L'analyse au moyen du modèle mixte peut être effectuée par exemple au moyen de l'instruction PROC MIXED du logiciel SAS. Dans le cas de la deuxième analyse, deux options sont importantes. D'abord, il faut calculer les degrés de liberté selon la méthode exposée dans la référence [19], ce qui, dans le logiciel SAS, se fait en choisissant l'option `ddfm=KR` dans l'instruction `model`. Ensuite, la structure de la covariance non structurée inter-sujets hétérogène (UN-H) doit être ajustée [23], en employant les options type=UN group=groupingvar dans les instructions répétées, où groupingvar est le nom de la variable qui contient la classification du groupe.

Références

[1] Algina, J. (1997). Generalization of Improved General Approximation tests to split-plot designs with multiple between-subjects factors and/or multiple within-subjects factors. *British Journal of Mathematical and Statistical Psychology, 50,(2),* 243-252.

[2] Algina, J. et Keselman, H. J. (1997). Detecting repeated measures effects with univariate and multivariate statistics. *Psychological Methods, 2*(2), 208-218.

[3] Altman, D. G. et Bland, J. M. (1995). Statistics notes: Absence of evidence is not evidence of absence. *British Medical Journal, 311*(7003), 485-485.

[4] Arnau, J., Bendayan, R., Blanca, M. J. et Bono, R. (2013). The effect of skewness and kurtosis on the robustness of linear mixed models. *Behavior Research Methods, 45*(3), 873-879. doi: 10.3758/s13428-012-0306-x.

[5] Berkovits, I., Hancock, G. R. et Nevitt, J. (2000). Bootstrap resampling approaches for repeated measure designs: relative robustness to sphericity and normality violations. *Educational and Psychological Measurement, 60*(6), 877-892.

[6] Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2ème éd.). Hillsdale, N.J.: L. Erlbaum Associates.

[7] Conover, W. J. (1999). Practical nonparametric statistics (3ème éd.). New York: Wiley.

[8] Cramér, H. (1946). Mathematical methods of statistics. Princeton: Princeton University Press.

[9] DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological Methods, 2*(3), 292-307. doi: 10.1037//1082-989x.2.3.292.

[10] Doornik, J. A. et Hansen, H. (2008). An omnibus test for univariate and multivariate normality. *Oxford Bulletin of Economics and Statistics, 70,(s1),* 927-939. doi: 10.1111/j.1468-0084.2008.00537.x.

[11] Glass, G. V., Peckham, P. D. et Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying fixed effects analyses of variance and covariance. *Review of Educational Research, 42*(3), 237-288. doi: 10.3102/00346543042003237.

[12] Greenhouse, S. W. et Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika, 24*(2), 95-112.

[13] Harwell, M. R., Rubinstein, E. N., Hayes, W. S. et Olds, C. C. (1992). Summarizing Monte-Carlo results in methodological research: The one-factor and two-factor fixed effects ANOVA cases. *Journal of Educational and Behavioral Statistics, 17*(4), 315-339. doi: 10.3102/10769986017004315.

[14] Henze, N. et Zirkler, B. (1990). A class of invariant consistent tests for multivariate normality. *Communications in Statistics-Theory and Methods, 19*(10), 3595-3617. doi: 10.1080/03610929008830400.

[15] Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika, 75*(4), 800-802.

[16] Huynh, H. et Feldt, L. S. (1970). Conditions under which mean square ratios in repeated measurements designs have exact *F*-distributions. *Journal of the American Statistical Association, 65*(332), 1582-1589.

[17] Huynh, H. et Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational aand Behavioral Statistics, 1*(1), 69-82. doi: http://dx..org/10.2307/1164736.

[18] Jensen, D. R. (1982). Efficiency and robustness in the use of repeated measurements. *Biometrics, 38*(3), 813-825. doi: 10.2307/2530060.

[19] Kenward, M. G. et Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics, 53*(3), 983-997.

[20] Keselman, H. J. (1994). Stepwise and simultaneous multiple comparison procedures of repeated measures' means. *Journal of Educational and Behavioral Statistics, 19*(2), 127-162.

[21] Keselman, H. J., Algina, J. et Kowalchuk, R. K. (2001). The analysis of repeated measures designs: A review. *British Journal of Mathematical & Statistical Psychology, 54, (1),* 1-20.

[22] Keselman, H. J., Kowalchuk, R. K., Algina, J., Lix, L. M. et Wilcox, R. R. (2000). Testing treatment effects in repeated measures designs: Trimmed means and bootstrapping. *British Journal of Mathematical & Statistical Psychology, 53,(2),* 175-191.

[23] Kowalchuk, R. K., Keselman, H. J., Algina, J. et Wolfinger, R. D. (2004). The analysis of repeated measurements with mixed-model adjusted *F* tests. *Educational and Psychological Measurement, 64*(2), 224-242. doi: 10.1177/0013164403260196.

[24] Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D. et Schabenberger, O. (2006). *SAS for mixed models* (2nd ed.). Cary, N.C.: SAS Institute, Inc.

[25] Lix, L. M., Keselman, J. C. et Keselman, H. J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. *Review of Educational Research, 66*(4), 579-619. doi: 10.2307/1170654.

[26] Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika, 57*(3), 519-530. doi: 10.2307/2334770.

[27] Maxwell, S. E. (1980). Pairwise multiple comparisons in repeated measures designs. *Journal of Educational and Behavioral Statistics, 5*(3), 269-287. doi: 10.3102/10769986005003269.

[28] Maxwell, S. E. et Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, N.J.: Lawrence Erlbaum Associates.

[29] Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 156-166.

[30] Oberfeld, D. et Franke, T. (2013). Evaluating the robustness of repeated measures analyses: The case of small sample sizes and non-normal data. *Behavior Research Methods, 45*(3), 792-812. doi: http:/dx.doi.org/10.3758/s13428-012-0281-2.

[31] Olejnik, S. et Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods, 8*(4), 434-447. doi: 10.1037/1082-989x.8.4.434.

[32] Rasmussen, J. L. (1987). Parametric and Bootstrap Approaches to Repeated Measures Designs. *Behavior Research Methods Instruments & Computers, 19*(4), 357-360.

[33] Rouanet, H. et Lépine, D. (1970). Comparison between treatments in a repeated-measurement design: ANOVA and multivariate methods. *British Journal of Mathematical and Statistical Psychology, 23*(2), 147-163.

[34] Royston, J. P. (1983). Some techniques for assessing multivariate normality based on the Shapiro-Wilk-W. *Applied Statistics-Journal of the Royal Statistical Society Series C, 32*(2), 121-133. doi: 10.2307/2347291.

[35] Schmider, E., Ziegler, M., Danay, E., Beyer, L. et Bühner, M. (2010). Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption. *Methodology-European Journal of Research Methods for the Behavioral and Social Sciences, 6*(4), 147-151. doi: 10.1027/1614-2241/a000016.

[36] St. Laurent, R. et Turk, P. (2013). The effects of misconceptions on the properties of Friedman's test. *Communications in Statistics-Simulation and Computation, 42*(7), 1596-1615. doi: 10.1080/03610918.2012.671874.

[37] Tukey, J. W. (1977). Exploratory data analysis. Reading, Mass.: Addison-Wesley Pub. Co.

[38] Seco, G. V., Izquierdo, M. C., García, M. P. F. et Díez, F. J. H. (2006). A comparison of the bootstrap-F, improved general approximation, and Brown-Forsythe multivariate approaches in a mixed repeated measures design. *Educational and Psychological Measurement, 66*(1), 35-62.

[39] Wilcox, R. R., Keselman, H. J., Muska, J. et Cribbie, R. (2000). Repeated measures ANOVA: Some new results on comparing trimmed means and means. *British Journal of Mathematical & Statistical Psychology, 53*, 69-82.

Appendice 5
à l'Annexe 1
(à titre informatif)

Critères permettant d'obtenir un comportement optimal des repères

Un bon repère doit être conçu de façon à satisfaire le mieux possible aux caractéristiques essentielles énoncées ci-après.

Un repère dont le comportement est optimal doit permettre:

1) de produire des données qui ne présentent pas de changements substantiels, en ce qui concerne le classement des systèmes testés, par rapport aux données recueillies au moyen des spécifications applicables aux repères de la Recommandation UIT-R BS.1534;

2) d'être associé aux notes de l'auditeur qui s'étendent sur une plage plus large de l'échelle de notation pour les systèmes testés, par rapport aux données recueillies pour les systèmes testés au moyen des spécifications applicables aux repères de la Recommandation UIT-R BS.1534;

3) d'être perçu par les auditeurs comme ressemblant plus aux systèmes testés que les repères décrits par les spécifications de la Recommandation UIT-R BS.1534. Cela peut conduire à des temps plus longs d'évaluation des repères;

4) d'assurer une comparaison précise des systèmes testés situés au milieu de l'échelle;

5) de produire des notes pour les repères situés en bas et au milieu de l'échelle différant d'environ 20 à 30 points;

6) de produire des dégradations de la qualité, s'agissant des repères qui ne dépendent que peu du contenu.