RECOMMANDATION UIT-R BS.1534

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

(Question UIT-R 220/10)

(2001)

L'Assemblée des radiocommunications de l'UIT,

considérant

a) que l'on a élaboré dans les Recommandations UIT-R BS.1116, UIT-R BS.1284, UIT-R BT.500, UIT-R BT.710 et UIT-R BT.811, ainsi que dans les Recommandations UIT-T P.800, UIT-T P.810 et UIT-T P.830, des méthodes pour l'évaluation de la qualité subjective des systèmes audio, vidéo et vocaux;

b) que de nouveaux types de services de distribution tels que la diffusion audio sur Internet ou via des lecteurs statiques à semi-conducteurs, les services numériques par satellite, les systèmes numériques à ondes courtes ou ondes moyennes, ou les applications multimédias 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 qu'elle 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'on s'attache essentiellement, dans les Recommandations UIT-T P.800, UIT-T P.810 et UIT-T P.830, aux signaux vocaux dans un environnement téléphonique, ce qui se révèle insuffisant pour l'évaluation des signaux audio dans un environnement de radiodiffusion;

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

g) que de nouveaux services multimédias peuvent nécessiter une évaluation conjointe des qualités audio et vidéo,

recommande

1 d'utiliser les procédures de test et d'évaluation exposées dans l'Annexe 1 de la présente Recommandation pour l'évaluation subjective de la qualité audio intermédiaire.

ANNEXE 1

1 Introduction

On décrit dans la présente Recommandation une nouvelle méthode d'évaluation subjective de la qualité audio intermédiaire. Cette méthode reflète plusieurs aspects de la Recommandation UIT-R BS.1116 et fait appel à la même échelle d'évaluation que celle utilisée pour l'évaluation de la qualité d'image (voir la Recommandation UIT-R BT. 500).

Cette méthode, appelée «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 qu'elle conduit à des résultats exacts et fiables [UER, 2000a; Soulodre et Lavoie, 1999; UER, 2000b].

La présente Recommandation comprend les Sections et Appendice suivants:

Section 1: Introduction

Section 2: Portée, intérêt du test et but de la nouvelle méthode

Section 3: Conception des expériences

Section 4: Sélection des sujets

Section 5: Méthode de test

Section 6: Caractéristiques

Section 7: Données de test

Section 8: Conditions d'écoute

Section 9: Analyse statistique

Section 10: Rapport de test et présentation des résultats

Appendice 1: Instructions à donner aux auditeurs

2 Portée, intérêt du test 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. Il s'agit de méthodes largement décrites et avérées permettant d'évaluer la qualité audio aux deux extrêmes de la gamme.

On utilise la Recommandation UIT-R BS.1116 – Méthodes d'évaluation subjective des dégradations faibles dans les systèmes audio y compris les systèmes sonores multivoies, pour évaluer les systèmes audio de haute qualité qui présentent de faibles dégradations. Il existe toutefois des applications pour lesquelles des sons audio de moindre qualité sont acceptables ou inévitables. Le rapide développement de l'utilisation de l'Internet pour la diffusion et la radiodiffusion de données audio, avec des débits de données limités, ont conduit à un compromis quant à la qualité audio. On peut citer d'autres applications associées à une qualité audio intermédiaire, telles que: la modulation d'amplitude numérique (par exemple, la radio mondiale numérique (DRM, *digital radio mondiale*)), la radiodiffusion numérique par satellite, les circuits de commentaire en radio et télévision, 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 pour l'évaluation de ces systèmes audio de moindre qualité [Soulodre et Lavoie, 1999], parce qu'elle ne permet une bonne discrimination entre des niveaux de qualité semblables en bas de l'échelle d'évaluation.

Dans la Recommandation UIT-R BS.1284 ne figurent que des méthodes qui ne concernent que la haute qualité audio 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 de projet B/AIM de l'Union Européenne de Radio-Télévision (UER) a effectué des expériences s'appuyant sur ces méthodes de l'UIT-T en utilisant

des données audio typiques d'un environnement de radiodiffusion. Aucune de ces méthodes ne répond au «cahier des charges» d'une échelle absolue, (comparaison avec un signal de référence et petits intervalles de confiance avec un nombre raisonnable de sujets durant une même période). L'évaluation des signaux audio dans un environnement de radiodiffusion ne peut donc pas être effectuée correctement en utilisant l'une de ces méthodes.

La nouvelle méthode de test décrite dans la présente Recommandation doit permettre de fournir une mesure fiable et reproductible pour des systèmes dont la qualité audio se situerait en principe dans la moitié inférieure de l'échelle de dégradation utilisée dans la Recommandation UIT-R BS.116 [UER, 2000a; Soulodre et Lavoie, 1999; UER, 2000b]. Dans la méthode de test MUSHRA, on utilise un signal de référence de haute qualité et les systèmes provoquent généralement des dégradations significatives. D'autres méthodes de test doivent être utilisées si les systèmes testés sont susceptibles d'améliorer la qualité subjective d'un signal.

3 Conception des expériences

On utilise de nombreux différents types de stratégie de recherche pour rassembler des informations fiables dans un domaine présentant un intérêt scientifique. Il convient d'utiliser les méthodes expérimentales les plus formelles pour évaluer de manière subjective les dégradations des systèmes audio. Les expériences subjectives se caractérisent d'abord par une maîtrise réelle des conditions expérimentales, puis par le rassemblement et l'analyse de données statistiques émanant d'auditeurs. Il est nécessaire de procéder à une conception et à une planification soignées des expériences pour faire en sorte que des facteurs non maîtrisés pouvant créer des ambiguïtés dans les résultats des tests, soient minimisés. Par exemple, si la séquence réelle des éléments sonores est la même pour tous les sujets participant à un test d'écoute, on ne peut pas savoir si les jugements portés par les participants sont motivés par l'écoute de cette séquence plutôt que par les différents niveaux de dégradation présentés. Ainsi, les conditions de test doivent être telles qu'elles ne permettent de mettre en évidence que les seuls effets des facteurs indépendants.

Dans les situations où l'on peut supposer que les dégradations potentielles et les autres caractéristiques seront réparties de manière homogène tout au long du test d'écoute, il conviendrait peut-être de rendre la présentation des conditions de test tout à fait aléatoire. Si l'on s'attend à une répartition hétérogène, il faut en tenir compte dans la présentation des conditions de test. Par exemple, si les éléments à évaluer présentent des niveaux de difficulté variables, il faut présenter les stimuli de manière aléatoire, tant au cours d'une même séance que d'une séance à l'autre.

Il convient de concevoir les tests d'écoute de manière à ne pas fatiguer les participants au point de réduire la précision de leur jugement. Sauf lorsque la relation entre le son et la vision est importante, il est préférable que l'évaluation des systèmes audio s'effectue sans présentation d'images associées. La présence de conditions de contrôle appropriées doit être un sujet de préoccupation essentiel. Il s'agit généralement d'insérer des éléments audio non dégradés sans que les participants ne puissent les déceler. Les différences entre les jugements relatifs à ces stimuli de contrôle et ceux relatifs aux stimuli potentiellement dégradés permettent de déduire si l'échelle de notation permet une bonne évaluation des dégradations.

On décrira plus loin certaines de ces considérations. Il faut savoir que la conception des expériences, leur réalisation et leur analyse statistique sont des processus complexes, dont tous les détails ne peuvent être donnés dans une recommandation telle que celle-ci. Il est recommandé que des professionnels connaissant parfaitement la conception des expériences et les statistiques associées soient consultés ou présents au début de la planification du test d'écoute.

4 Sélection des sujets

Les données de test d'écoute destinées à évaluer les faibles dégradations du fonctionnement qualitatif des systèmes audio (Recommandation UIT-R BS.1116) doivent provenir de sujets ayant une expérience de ce type de test. Plus la qualité des systèmes à tester est élevée, plus il importe de disposer d'auditeurs expérimentés.

4.1 Critère de sélection des auditeurs

La méthode MUSHRA ne doit en principe pas être appliquée pour de faibles dégradations, il est cependant recommandé de faire appel à des auditeurs expérimentés. Ces personnes doivent avoir l'habitude d'écouter le son de manière critique. Les résultats ainsi obtenus seront plus fiables que s'il s'agissait d'auditeurs 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 de défaut après y avoir été fréquemment exposés.

Il existe parfois une raison motivant la mise en œuvre d'une technique de rejet avant (présélection) ou après (postsélection) le test réel. On peut utiliser dans certains cas les deux types de rejet. On entend ici par rejet un processus éliminant tous les jugements émanant d'un auditeur particulier.

Toute technique de rejet, si elle n'est pas analysée et appliquée avec soin, risque de conduire à un résultat biaisé. Il est donc extrêmement important que le rapport de test décrive clairement le critère appliqué pour toute élimination de données.

4.1.1 Présélection des auditeurs

Le groupe d'auditeurs doit être composé de sujets expérimentés, c'est-à-dire de personnes comprenant la méthode d'évaluation de la qualité subjective et ayant reçu la formation voulue. Ces auditeurs devront:

- avoir l'expérience d'une écoute critique des sons;
- disposer de capacités auditives normales (la norme ISO 389 faisant office de référence à cet effet).

La procédure d'entraînement pourrait servir de moyen de présélection.

L'argument majeur en faveur de l'introduction d'une technique de présélection réside dans l'accroissement de 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 Postsélection des auditeurs

On peut grosso modo classer les méthodes de postsélection en au moins deux catégories:

- l'une fondée sur la capacité du sujet à faire des évaluations répétées et cohérentes;
- l'autre repose sur les incohérences de l'évaluation individuelle comparée à la moyenne de toutes les évaluations d'un élément donné.

On recommande d'examiner la variation, l'écart, entre l'évaluation individuelle et la moyenne des notes données par un ensemble de participants.

On cherche ici à obtenir une évaluation suffisamment précise de la qualité des éléments testés.

Lorsque les évaluations d'un petit nombre de participants se situent à l'une des extrémités de l'échelle (excellent, mauvais) alors que la majorité des appréciations est regroupée en un autre point de cette échelle, on peut isoler voire éliminer ces observations aberrantes.

Puisque l'on teste la «qualité intermédiaire», un auditeur doit être à même d'identifier très facilement la version codée et donc donner une note correspondant au choix de la majorité des sujets. Les participants dont les appréciations se situent à l'extrémité supérieure de l'échelle sont vraisemblablement très peu critiques, alors que ceux qui donnent des notes «très basses» le sont vraisemblablement trop. Le rejet de ces valeurs extrêmes doit permettre une évaluation plus réaliste de la qualité.

On utilise essentiellement ces méthodes pour éliminer les participants incapables de faire les distinctions appropriées. La mise en œuvre d'une méthode de postsélection peut permettre de clarifier la «lecture» des résultats de test. Il convient cependant de rester prudent, eu égard aux différences de sensibilité des auditeurs aux divers défauts. En accroissant la taille du groupe d'auditeurs, l'incidence des notations d'un individu donné est réduite, ce qui diminue fortement la probabilité de devoir rejeter les données correspondantes.

4.2 Effectif d'un groupe d'écoute

Pour déterminer le nombre d'auditeurs qui convient, il faut pouvoir estimer la variance des notations des différents sujets et connaître la résolution requise pour l'expérience.

Lorsque les conditions d'un test d'écoute sont bien maîtrisées tant du point de vue technique que psychologique, l'expérience montre que les données d'une vingtaine de participants sont souvent suffisantes pour tirer du test les conclusions appropriées. Si l'analyse peut être conduite parallèlement au test, il n'est pas nécessaire d'ajouter des participants dès lors qu'un degré de précision statistique suffisant pour formuler des conclusions appropriées est atteint.

Lorsque, pour une raison quelconque, on ne maîtrise pas bien les conditions de l'expérience, il peut être nécessaire d'accroître le nombre d'auditeurs afin d'obtenir la résolution requise.

La taille du groupe de participants ne dépend pas seulement de la résolution voulue. 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. Ainsi, en accroissant la taille du groupe d'écoute, on peut dire que les résultats seraient valables pour un groupe d'auditeurs «plus général» et qu'ils seraient donc plus convaincants. Il peut également être nécessaire d'accroître l'effectif pour tenir compte de la probabilité de variation de la sensibilité des divers participants aux différents défauts.

5 Méthode de test

La méthode MUSHRA fait appel au programme d'origine non modifié, la largeur de bande totale constituant le signal de référence (également utilisé comme référence cachée) ainsi qu'au moins un repère caché. On peut utiliser d'autres repères bien définis (§ 5.1).

5.1 Description des signaux de test

Une séquence ne doit en principe pas durer plus de 20 s (il s'agit d'éviter de fatiguer les auditeurs et de réduire la durée totale du test d'écoute).

L'ensemble des signaux traités comprend tous les signaux testés ainsi qu'au moins un signal supplémentaire (repère) qui est le signal non traité obtenu après filtrage passe-bas. La largeur de bande de ce signal supplémentaire doit être de 3,5 kHz. Suivant le contexte du test, on peut utiliser d'autres repères. On peut également utiliser d'autres types de repère présentant des types de dégradations semblables à celles des systèmes testés, par exemple:

– limitation de la largeur de bande à 7 kHz ou 10 kHz;

– image stéréo réduite;

– bruit supplémentaire;

– pertes de signal;

– pertes de paquets;

– divers.

NOTE 1 – Les largeurs de bande des repères correspondent aux 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), visés respectivement aux Recommandations UIT-T G.711, UIT-T G.712, UIT-T G.722 et UIT-T J.21. Les caractéristiques du filtre passe-bas de 3,5 kHz doivent être les suivantes:

f_c = 3,5 kHz,

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

affaiblissement minimum à 4 kHz = 25 dB,

affaiblissement minimal à 4,5 kHz = 50 dB.

Les repères supplémentaires doivent permettre de comparer les systèmes testés à des niveaux de qualité audio bien connus et ne doivent pas être utilisés pour réévaluer les résultats entre différents tests.

5.2 Phase de familiarisation

Pour obtenir des résultats fiables, il est nécessaire de former les participants lors de séances spéciales préalables au test. On a constaté que cette familiarisation était importante pour obtenir des résultats fiables. Il faut à cette occasion que les participants soient à tout le moins confrontés à toute la gamme et à tous les types de dégradations ainsi qu'à l'ensemble de tous les signaux mis en œuvre durant le test. Plusieurs méthodes peuvent être utilisées à cet effet (simple système de lecture de cassette audio, système interactif assisté par ordinateur, etc.). Des instructions correspondantes figurent dans l'Appendice 1.

5.3 Présentation des stimuli

La méthode MUSHRA utilisée ici est une méthode de test en double aveugle à stimuli multiples avec une référence cachée et un ou plusieurs repère(s) caché(s), alors que l'on utilise dans la Recommandation UIT-R BS.1116 une méthode de test «en double aveugle à triple stimuli avec une référence cachée». On considère que la méthode MUSHRA est plus appropriée pour l'évaluation des dégradations moyennes et fortes [Soulodre et Lavoie, 1999].

Lors d'un test portant sur de faibles dégradations, la difficulté, pour le sujet, est de détecter tous les défauts susceptibles d'être présents dans le signal. Il est nécessaire dans ce cas de disposer d'un signal de référence caché pour permettre à l'expérimentateur d'évaluer la capacité du sujet à bien

détecter ces défauts. En revanche, dans un test comprenant des dégradations moyennes et fortes, le participant n'a aucune difficulté à détecter les défauts, ce qui rend donc inutile le recours à une référence cachée. La difficulté se présente en fait lorsque le sujet doit noter les effets perturbateurs relatifs des différents défauts. Il doit alors relativiser ses «préférences» pour tel ou tel type de défaut.

L'utilisation d'une référence de haute qualité conduit à un problème intéressant. Puisque cette nouvelle méthode doit être utilisée pour évaluer les détériorations moyennes et fortes, on s'attend à ce que l'écart de perception entre le signal de référence et les éléments de test soit relativement grand. En revanche, l'écart de perception entre les éléments de test relevant de systèmes différents peut être très faible. Ainsi, si l'on utilise une méthode de test multicritères (comme celle utilisée dans la Recommandation UIT-R BS.1116), il peut être très difficile pour les participants de distinguer précisément les différents signaux détériorés. Par exemple, dans un test de comparaison directe couplée, les auditeurs peuvent tous dire que le Système A est meilleur que le Système B. Cependant, si l'on compare uniquement chaque système au signal de référence (les Systèmes A et B ne sont plus directement comparés entre eux), les différences entre les deux systèmes peuvent disparaître.

Pour surmonter cette difficulté dans la méthode de test MUSHRA, l'auditeur peut à volonté commuter le signal de référence et l'un quelconque des systèmes testés, généralement en utilisant un système de lecture assisté par ordinateur (mais on peut aussi utiliser des lecteurs multiCD ou multicassettes). On présente à l'auditeur une séquence d'essais. Pour chaque essai, on présente au sujet la version de référence ainsi que toutes les versions du signal de test traitées par les systèmes testés. Par exemple, si un test porte sur 8 systèmes audio, le sujet a la possibilité de commuter instantanément 11 signaux différents (un signal de référence + 8 signaux dégradés + 1 signal de référence caché + 1 signal de repère caché).

Le sujet pouvant directement comparer les signaux dégradés, cette méthode permet de bénéficier des avantages d'un test de comparaison par paires complet puisque le sujet peut plus facilement détecter les différences entre les signaux dégradés avant de les noter en conséquence. Cette caractéristique permet d'obtenir un degré de résolution élevé pour les notes attribuées aux systèmes. Il est important de remarquer toutefois que les participants attribueront leur note à un système donné après avoir comparé ce système au signal de référence, ainsi qu'aux autres signaux relatifs à chaque essai.

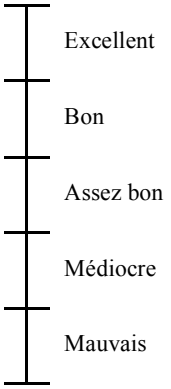
On recommande de ne pas faire intervenir plus de 15 signaux par essai (par exemple, 12 systèmes testés, 1 signal de référence connue, 1 signal repère caché et un signal de référence cachée).

Dans le test décrit dans la Recommandation UIT-R BS.1116, les participants commencent généralement un essai par un processus de détection, suivi par un processus de notation. L'expérience de tests conduits selon la méthode MUSHRA montre que les participants commencent généralement une séance avec une estimation approximative de la qualité. Suit un processus de tri ou de classement, qui précède le processus de notation. Le processus de classification étant effectué de manière directe, il est probable que les résultats relatifs à la qualité audio intermédiaire seront plus cohérents et fiables que si l'on avait utilisé la méthode de la Recommandation UIT-R BS.1116.

5.4 Processus de graduation

On demande aux participants de noter les stimuli suivant l'échelle de qualité continue (CQS, *continuous quality scale*). Il s'agit d'échelles graphiques identiques (longues généralement d'au moins 10 cm) divisées en cinq intervalles égaux avec les mentions indiquées de haut en bas dans la Fig. 1.

FIGURE 1

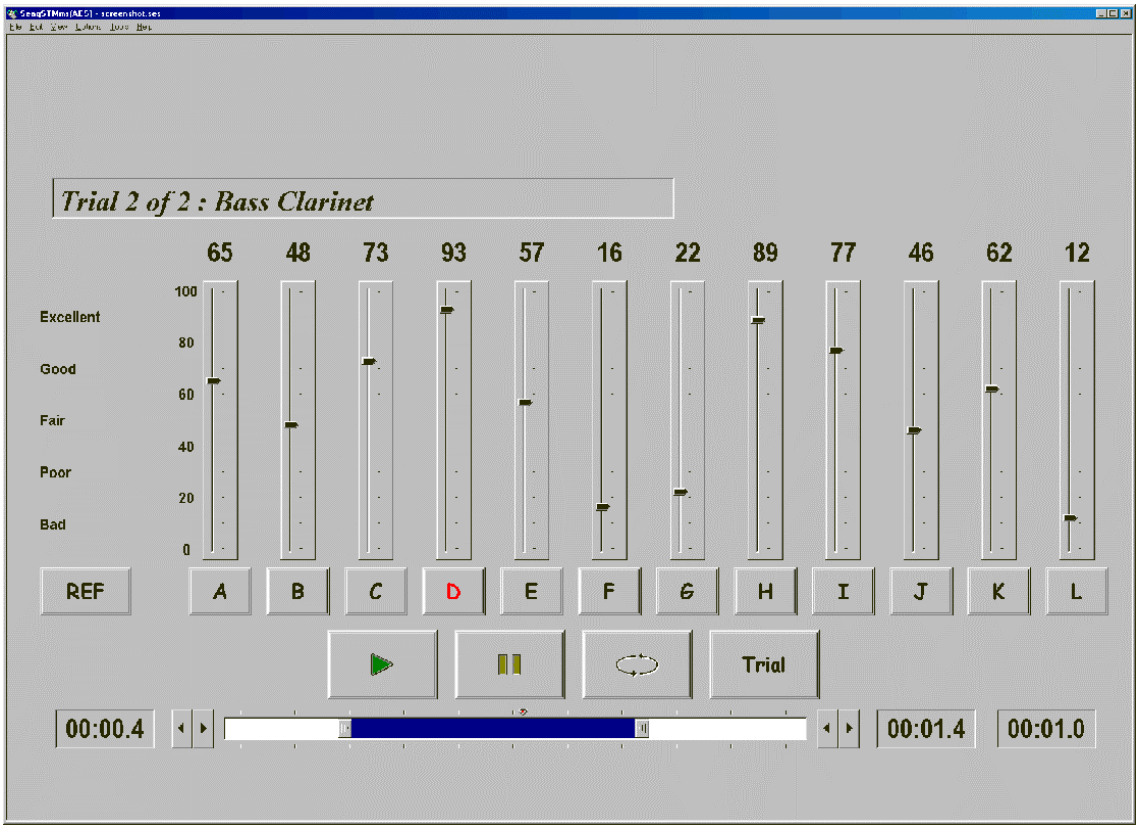


On utilise également cette échelle pour l'évaluation de la qualité d'image (Recommandation UIT-R BT.500 – Méthodologie d'évaluation subjective de la qualité des images de télévision).

L'auditeur enregistre son évaluation de la qualité de manière adéquate, à l'aide par exemple de curseurs dans une fenêtre électronique (voir la Fig. 2), ou en utilisant un crayon et une graduation sur papier. On demande à l'auditeur d'évaluer la qualité de tous les stimuli en fonction de l'échelle CQS à cinq intervalles.

FIGURE 2

Exemple d'affichage informatique utilisé pour un test MUSHRA



La méthode MUSHRA présente l'avantage, contrairement aux méthodes figurant dans la Recommandation UIT-R BS.1116, de présenter tous les stimuli en même temps, ce qui permet au participant de faire directement toutes les comparaisons de son choix. La cohérence des résultats s'en trouve accrue, ce qui conduit à des intervalles de confiance plus petits. La durée du test, lorsqu'on utilise la méthode MUSHRA, peut être sensiblement réduite par rapport à la méthode exposée dans la Recommandation UIT-R BS.1116.

6 Caractéristiques

On trouvera ci-dessous la liste des caractéristiques spécifiques des évaluations d'applications monophoniques, stéréophoniques et multivoies. Il est préférable d'évaluer dans chaque cas la caractéristique «qualité audio de base». Les expérimentateurs peuvent choisir de définir et d'évaluer d'autres caractéristiques.

Il convient de ne noter qu'une caractéristique par séance. Si l'on demande aux participants d'évaluer plus d'une caractéristique par essai, les participants peuvent se sentir perdus et/ou déconcertés, ayant à répondre à de multiples questions relatives à un stimulus donné. Le risque est alors de voir apparaître des notations peu fiables pour toutes ces questions.

6.1 Système monophonique

Qualité audio de base: on se réfère à cette caractéristique unique et globale pour évaluer toutes les différences décelées entre la référence et l'objet du test.

6.2 Système stéréophonique

Qualité audio de base: on utilise cette caractéristique unique et globale pour évaluer toutes les différences décelées entre la référence et l'objet du test. La caractéristique supplémentaire suivante peut être intéressante:

Qualité d'image stéréophonique: cette caractéristique est associée à la différence entre la référence et l'objet en terme d'emplacement des images sonores, d'impression de profondeur et de présence de l'événement audio. Bien que certaines études aient montré que la qualité d'image stéréophonique puisse être dégradée, cette question n'a pas encore fait l'objet d'études suffisantes permettant de justifier la pertinence d'une notation de la qualité d'image stéréophonique séparée et distincte de la qualité audio de base.

NOTE 1 – Jusqu'en 1993, la plupart des études d'évaluation subjective des faibles dégradations pour des systèmes stéréophoniques portaient uniquement sur la qualité audio de base. Ainsi, la qualité d'image stéréophonique était implicitement ou explicitement incluse dans la qualité audio de base, considérée dans ces études comme une caractéristique globale.

6.3 Système multivoies

Qualité audio de base: on utilise cette caractéristique unique et globale pour toutes les différences décelées entre la référence et l'objet du test.

Les caractéristiques supplémentaires suivantes peuvent être intéressantes:

Qualité frontale de l'image: cette caractéristique est associée à la localisation des sources sonores frontales. Elle comprend la qualité d'image stéréophonique et les pertes de définition.

Qualité d'impression ambiophonique: cette caractéristique est associée à une impression d'espace, à l'ambiance ou à des effets d'ambiophonie directionnels particuliers.

7 Données de test

Il convient d'utiliser des données de test critiques représentatives du programme de radiodiffusion typique de l'application considérée, afin de faire apparaître les différences entre les systèmes testés. Les données critiques sont celles qui mettent en évidence les caractéristiques des systèmes testés. Il n'existe pas de données de programmes universelles pouvant être utilisées pour évaluer tous les systèmes dans toutes les conditions. Il faut au contraire que des données de programmes adaptées soient recherchées spécifiquement pour chaque système testé et chaque type d'expérience. La recherche de données adaptées prend en général beaucoup de temps; cependant, si on ne parvient pas à trouver des données critiques adaptées à chaque système, les expériences ne permettront pas de faire apparaître des différences entre les systèmes et ne permettront pas de dégager des conclusions.

On doit pouvoir prouver de manière empirique et statistique que tout échec de différenciation entre des systèmes n'est pas dû à des conditions expérimentales trop peu sensibles qui peuvent être occasionnées par un mauvais choix de données audio, ou par un autre aspect insuffisant de l'expérience. Dans le cas contraire, on ne peut accepter cette «absence» de différenciation.

Dans cette recherche de données critiques, il faut accepter tout stimulus qui pourrait se présenter dans un programme radiodiffusé. Les signaux artificiels délibérément conçus pour perturber un système particulier ne doivent pas être pris en compte. Le contenu artistique ou intellectuel d'une séquence de programmes ne doit jamais présenter un degré d'intérêt, de désagrément ou d'ennui tel qu'il détourne l'attention de l'auditeur et l'empêche de se concentrer pour détecter d'éventuelles erreurs. Il convient de tenir compte de la fréquence d'apparition probable de chaque type d'élément de programmes dans les émissions de radiodiffusion réelles. Il faut toutefois garder à l'esprit que les programmes de radiodiffusion sont susceptibles d'être différents à l'avenir du fait de l'évolution des styles et préférences musicaux.

Lorsque l'on choisit les éléments de programme, il est important de définir avec précision les caractéristiques à évaluer. Il convient de confier la responsabilité du choix des éléments à un groupe d'auditeurs compétents possédant une connaissance de base des dégradations probables. Les auditeurs doivent au départ pouvoir s'appuyer sur une très large gamme d'éléments, pouvant être élargie par l'ajout d'enregistrements spéciaux.

Dans la préparation du test subjectif, il est nécessaire que le groupe d'auditeurs qualifiés règle de manière subjective le volume de chaque extrait avant de l'enregistrer sur le support de test. On pourra ensuite se servir de ce support pour régler de façon fixe le gain de tous les éléments de programme.

Pour toutes les séquences de test, le groupe d'auditeurs qualifiés doit parvenir à un consensus quant aux niveaux sonores relatifs des différents extraits de test. De plus, les experts doivent parvenir à un accord concernant le niveau de pression acoustique absolu reproduit par rapport au niveau d'alignement pour l'ensemble d'une séquence.

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 de sortie au niveau d'alignement d'entrée requis par le canal de reproduction, conformément à la Recommandation R 68 de l'UER (voir la Recommandation UIT-R BS.1116, paragraphe 8.4.1). La salve de tonalité n'est utile que pour l'alignement: elle ne doit pas être réutilisée durant le test. Il faut contrôler le signal radiophonique de façon que les crêtes d'amplitude ne dépassent que rarement l'amplitude crête du signal maximal admissible défini dans la Recommandation UIT-R BS.645 (onde sinusoïdale dépassant de 9 dB le niveau d'alignement).

Le nombre d'extraits que l'on peut inclure dans un test est variable: 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. Un extrait audio durera généralement de

10 à 20 s. Etant donné la difficulté de la tâche, il faut que les systèmes à tester soient disponibles. On ne peut réussir le processus de sélection que si l'on définit un calendrier approprié.

Il convient de tester la qualité de fonctionnement d'un système multivoies en reproduction à deux voies en utilisant un mixage réducteur de référence. Bien que le recours à un mixage réducteur puisse être jugé restrictif dans certaines conditions, il s'agit sans conteste de l'option la plus raisonnable à utiliser à long terme, pour 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 appropriés pour l'évaluation critique de la qualité de fonctionnement du mixage réducteur de référence à deux voies doit être fondée sur la reproduction d'éléments traités par mixage réducteur à deux voies.

8 Conditions d'écoute

On définit dans la Recommandation UIT-R BS.1116 des méthodes pour l'évaluation subjective de faibles dégradations dans les systèmes audio, dont les systèmes sonores multivoies. Il convient d'utiliser les conditions d'écoute décrites dans les § 7 et 8 de la Recommandation UIT-R BS.1116 pour évaluer les systèmes audio de qualité intermédiaire.

On peut utiliser des casques d'écoute ou des haut-parleurs. Leur utilisation conjointe dans une même séance de test est interdite: tous les sujets 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 canal de reproduction (par exemple un amplificateur de puissance et son haut-parleur associé), il faudra régler le gain de l'amplificateur au niveau de pression acoustique de référence (pondération CEI/A, lente):

$$L_{ref} = 85 - 10 \log n \pm 0{,}25 \qquad \text{dBA}$$

où n est le nombre de canaux de reproduction présents dans l'ensemble du dispositif.

On autorise les auditeurs à régler individuellement leur niveau d'écoute au cours d'une séance dans une limite 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 du test de manière telle que les auditeurs n'aient en principe pas besoin de procéder à des réglages individuels pour chaque élément.

Les réglages de niveau au sein d'un élément doivent être interdits.

9 Analyse statistique

Les mesures de longueur sur les feuilles de notation relatives aux évaluations de chaque condition de test sont converties linéairement en notes normalisées dans une gamme de 0 à 100, où 0 correspond au bas de l'échelle (mauvaise qualité). On calcule ensuite les notes absolues suivant la méthode développée ci-après.

Le calcul des moyennes des notes normalisées de tous les auditeurs restant après la postsélection conduira aux notes subjectives moyennes.

La première étape de l'analyse des résultats correspond au calcul de la notation moyenne $\overline{u}_{jk}$ pour chacune des présentations:

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

où:

u_i: note de l'observateur i pour une condition de test j et une séquence audio k

N: nombre d'observateurs

De la même manière, on pourrait calculer les notes moyennes globales, $\overline{u}_j$ et $\overline{u}_k$, pour chaque condition de test et chaque séquence de test.

Lorsque l'on présente les résultats d'un test, toutes les notes moyennes doivent être associées à un intervalle de confiance déduit de l'écart type et de la taille de chaque échantillon.

On propose d'utiliser l'intervalle de confiance à 95% donné par:

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

où

$$\delta_{jk} = t_{0,05}\,\frac{S_{jk}}{\sqrt{N}} \tag{2}$$

et $t_{0,05}$ correspond à la valeur de t pour un intervalle de confiance de 95%.

L'écart type pour chaque présentation, S_{jk}, est donné par:

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

La valeur absolue de la différence entre la note moyenne expérimentale et la note moyenne réelle (pour un très grand nombre d'observateurs) est inférieure, avec une probabilité de 95%, à la valeur de l'intervalle de confiance à 95%, à condition que la distribution des notes individuelles satisfasse à certains critères.

On pourrait de la même manière calculer l'écart type S_j pour chaque condition de test. On note toutefois que, lorsqu'on utilise un petit nombre de séquences de test, cet écart type dépend davantage des différences entre les séquences de test utilisées que des différences d'appréciation entre les évaluateurs.

L'expérience a montré que les notes correspondant à différentes séquences de test dépendent du caractère critique des éléments de test. On peut parvenir à une meilleure compréhension de la qualité de fonctionnement du système en présentant séparément les résultats des différentes séquences de test, plutôt qu'en indiquant simplement les moyennes globales relatives à l'ensemble des séquences de test de l'évaluation.

10 Rapport de test et présentation des résultats

10.1 Observations générales

La présentation des résultats doit être conviviale: tout lecteur, qu'il soit novice ou expert, doit pouvoir obtenir des informations pertinentes. Tout lecteur souhaite d'abord visualiser l'ensemble des résultats de l'expérience, sous forme graphique de préférence. On peut accompagner cette présentation d'informations quantitatives plus détaillées, même si le détail des analyses numériques doit se trouver en appendice.

10.2 Contenu du rapport de test

Le rapport de test doit indiquer aussi clairement que possible la logique de l'étude, les méthodes utilisées et les conclusions dégagées. Les informations doivent être suffisamment détaillées pour qu'une personne compétente puisse en principe reproduire cette étude afin d'en vérifier les résultats de manière empirique. Il n'est pas nécessaire toutefois que le rapport contienne l'ensemble des résultats. Un lecteur averti doit être à même de comprendre et de développer une opinion critique sur les points les plus importants du test, tels que les motifs fondamentaux de l'étude, les méthodes de conception et de réalisation de l'expérience, ainsi que l'analyse et les conclusions.

Il faut porter une attention particulière aux points suivants:

- représentation graphique des résultats;
- caractérisation et sélection des auditeurs (voir la Note 1);
- spécification et sélection des données de test;
- informations générales relatives au système utilisé pour traiter les éléments de test;
- détails de la configuration de test;
- détails matériels relatifs à l'environnement et au matériel d'écoute, dont 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 2);
- conception de l'expérience, formation, instructions, séquences de l'expérience, procédures de test, production de données;
- traitement des données, en particulier les détails relatifs aux statistiques déductives descriptives et analytiques;
- bases précises de toutes les conclusions tirées de l'étude.

NOTE 1 – Il est évident que les variations de capacités d'écoute entre les différents participants du groupe peuvent avoir une incidence sur les évaluations. Pour cerner encore davantage ce paramètre, les responsables de l'expérience sont invités à rendre compte autant que possible des caractéristiques du groupe d'auditeurs. L'âge et le sexe des participants peuvent par exemple constituer des éléments d'appréciation intéressants.

NOTE 2 – Puisqu'il est reconnu que les conditions d'écoute (la reproduction sonore par haut-parleurs plutôt que via des casques d'écoute par exemple) peut avoir une incidence sur les résultats des évaluations subjectives, les responsables de l'expérience sont invités à rendre compte explicitement des conditions d'écoute ainsi que du type d'équipement de reproduction sonore utilisé. Si on souhaite procéder à une analyse statistique combinée des différents types de transducteur, il convient de vérifier qu'une telle combinaison des résultats est possible (en utilisant le modèle ANOVA par exemple).

10.3 Présentation des résultats

Il faut indiquer pour chaque paramètre de test la moyenne et l'intervalle de confiance à 95% de la distribution statistique des notes d'évaluation.

Ces résultats doivent être accompagnés des informations suivantes:

– description des données de test;

– nombre d'auditeurs participant à l'évaluation;

– note moyenne globale pour tous les éléments de test utilisés dans l'expérience;

– notes moyennes et intervalles de confiance à 95% des seules observations formulées par les auditeurs ayant rempli les conditions de postsélection, (c'est-à-dire élimination de certains résultats conformément à la procédure décrite au § 4.1.2 (postsélection)).

On pourrait en outre présenter les résultats sous d'autres formes appropriées: histogrammes, médianes, etc.

10.4 Notes absolues

La présentation, pour les systèmes testés, des notes moyennes absolues, de la référence cachée ainsi que du repère donne un bon aperçu des résultats. Il convient toutefois de ne pas oublier que cela ne fournit aucun élément d'analyse statistique détaillée. Les observations n'étant pas indépendantes les unes des autres, l'analyse statistique de ces notes absolues ne conduira pas à des informations pertinentes et n'a donc pas d'utilité.

10.5 Niveau de signification et intervalle de confiance

La rapport de test doit fournir au lecteur des informations relatives à la nature statistique intrinsèque de toutes les données subjectives. Il convient d'indiquer les niveaux de signification ainsi que d'autres détails relatifs aux méthodes et résultats statistiques, pour aider le lecteur à comprendre les conclusions. On pourra par exemple faire figurer les intervalles de confiance ou les intervalles d'erreur sur les graphiques.

Il n'existe pas bien sûr un niveau de signification «correct». Toutefois, on retient en général la valeur de 0,05. Il est possible en principe d'utiliser un test à une ou à deux sorties en fonction de l'hypothèse testée.

Références bibliographiques

SOULODRE, G. A. et LAVOIE, M. C. [septembre 1999] Subjective evaluation of large and small impairments in audio codecs. AES 17th International Conference, Florence, p. 329-336.

UER [2000b] EBU Report on the subjective listening tests of some commercial internet audio codecs. Document BPN 029, juin.

UER [2000a] MUSHRA – Method for Subjective Listening Tests of Intermediate Audio Quality. Draft EBU Recommendation, B/AIM 022 (Rev.8)/BMC 607rev, janvier.

APPENDICE 1

À L'ANNEXE 1

Instructions à donner aux auditeurs

On trouvera ci-après un exemple de type d'instructions à donner ou à lire aux auditeurs pour les informer quant au déroulement du test.

1 Phase de familiarisation ou d'entraînement

Le but de la première étape des tests d'écoute est de vous familiariser à la procédure de test. Cette phase est appelée phase d'entraînement et précède la phase d'évaluation proprement dite.

Le but de la phase d'entraînement est de permettre à l'évaluateur que vous êtes d'atteindre les deux objectifs suivants:

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

– Partie B: apprendre à utiliser l'équipement de test et l'échelle de graduation.

Dans la Partie A de la phase d'entraînement, vous pourrez écouter plusieurs extraits sonores sélectionnés pour les tests afin d'illustrer toute la gamme des niveaux possibles de qualité. Les éléments sonores que vous écouterez seront de plus ou moins bonne qualité en fonction du débit binaire et des autres conditions choisies. La Fig. 3 présente l'interface utilisateur. Vous pouvez, en cliquant sur les différents boutons, écouter divers extraits sonores, dont les extraits de référence. Vous pouvez ainsi apprécier un ensemble de différents niveaux de qualité relatifs à divers éléments de programme. Ces extraits sont regroupés sur la base de conditions communes. Trois groupes ont été identifiés dans le cas de la Fig. 3, et chaque groupe comprend 4 signaux traités.

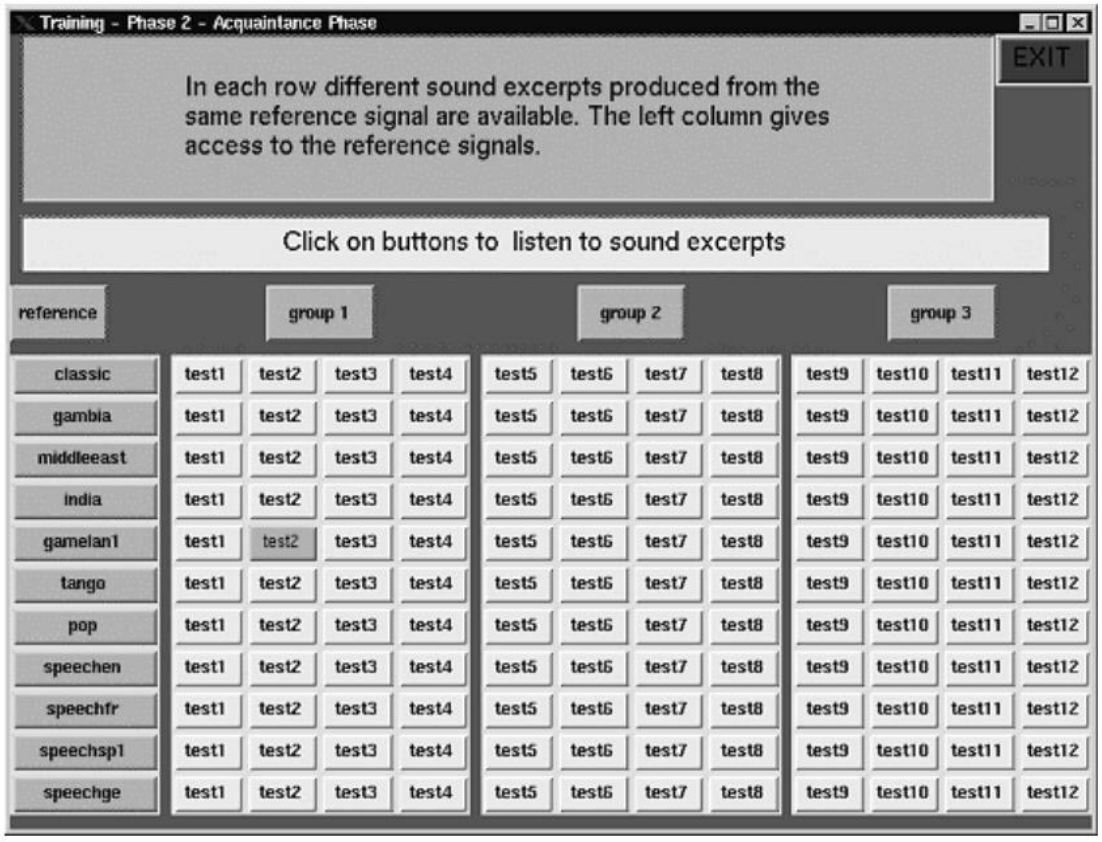
Dans la Partie B de la phase d'entraînement, vous apprendrez à utiliser les équipements de lecture et de notation qui vous seront nécessaires pour évaluer la qualité des extraits sonores.

2 Phase de notation en aveugle

Le but de la phase de notation en aveugle est de vous inviter à attribuer vos notes en utilisant l'échelle de qualité. Vos notes doivent traduire votre jugement subjectif du niveau de qualité de chacun des extraits qui vous sont proposés. Chaque jeu de test contient 11 signaux à noter. Chaque élément dure environ 10 à 20 s. Vous êtes invités à écouter l'extrait de référence ainsi que l'ensemble des conditions de test en cliquant sur les boutons correspondants. Vous pouvez écouter les signaux dans un ordre quelconque et autant de fois que vous le souhaitez. Veuillez utiliser le curseur associé à chaque signal pour indiquer votre opinion quant à sa qualité. Lorsque vous êtes satisfait des notes que vous avez attribuées à tous les signaux, cliquer sur le bouton «register scores» (*enregistrer les notes attribuées*) en bas de l'écran.

FIGURE 3

Exemple d'interface utilisateur pour la partie A de la phase d'entraînement



1534-03

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

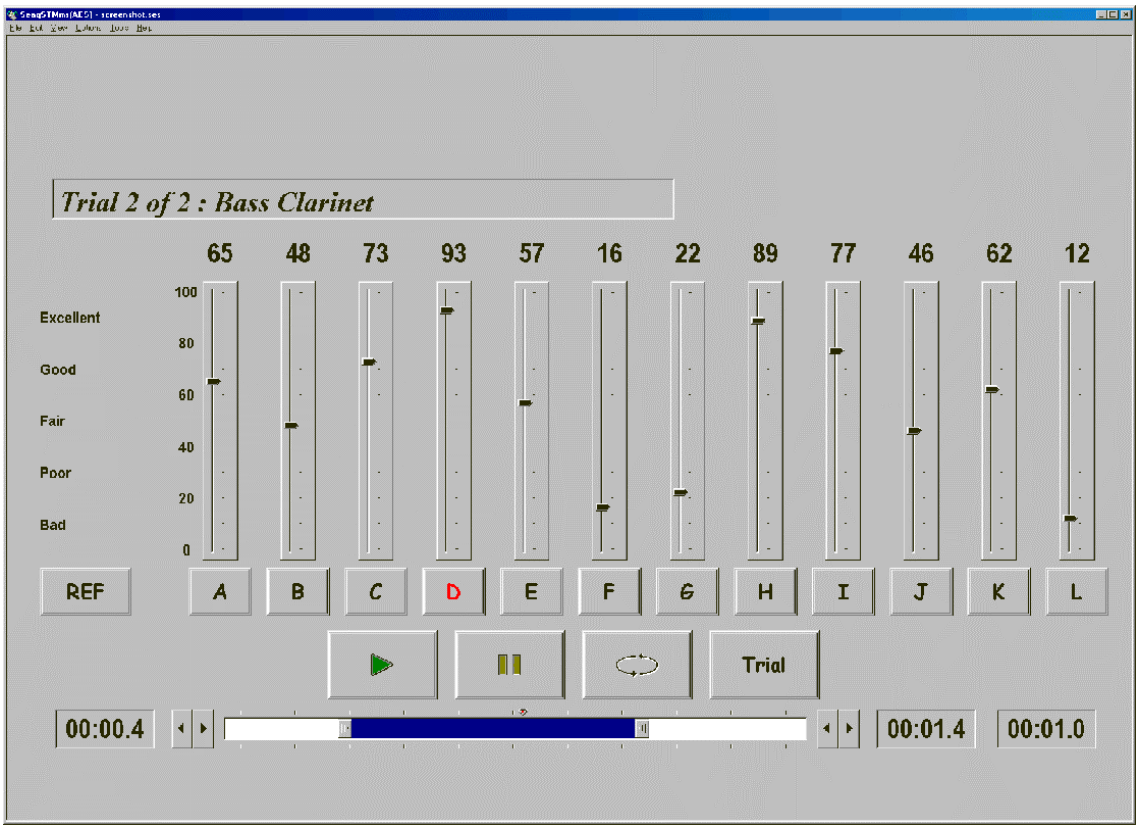
L'échelle de notation est continue entre les appréciations «excellent» et «mauvais». La note 0 correspond à l'extrémité inférieure de la catégorie «mauvais», alors que la note 100 correspond à l'extrémité supérieure de la catégorie «excellent».

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 du test présentant la qualité la plus faible. En revanche, vous devrez attribuer la note 100 à un ou plusieurs extraits sonores, car le signal de référence d'origine fait partie des extraits à évaluer.

Cette phase d'entraînement doit vous permettre d'apprendre à exprimer les dégradations perçues en notes d'évaluation. Vous ne devez jamais parler avec les autres auditeurs de vos appréciations personnelles durant la phase d'entraînement.

FIGURE 4

Exemple d'interface utilisateur mise en œuvre durant la phase de notation en aveugle



1534-04

Aucune note attribuée durant la phase d'entraînement ne sera prise en compte lorsque débuteront les tests réels.