Elaboration d'une méthode de test pour l'évaluation subjective de la qualité des sons spatialisés

Sarah Le Bagousse

THÈSE / UNIVERSITÉ DE BRETAGNE OCCIDENTALE

sous le sceau de l'Université européenne de Bretagne

pour obtenir le titre de

DOCTEUR DE L'UNIVERSITÉ DE BRETAGNE OCCIDENTALE

Mention : STIC
Spécialité : Acoustique

École Doctorale SICMA

présentée par

Sarah Le Bagousse

Préparée à Orange Labs, Rennes
Lab-STICC, Brest

Élaboration d'une méthode de test pour l'évaluation subjective de la qualité des sons spatialisés

Thèse soutenue le 29 avril 2014
devant le jury composé de :

Sylvain MARCHAND
Professeur, Université de Bretagne Occidentale / *directeur de thèse*

Mathieu PAQUIER
Maître de conférences, Université de Bretagne Occidentale / *encadrant de thèse*

Catherine COLOMES
Ingénieur, Orange Labs / *encadrant de thèse*

Etienne PARIZET
Professeur, INSA Lyon / *rapporteur*

Rozenn NICOL
Ingénieur, Orange Labs / *rapporteur*

Alexander RAAKE
Professeur, Deutsche Telekom / *examinateur*

A Pierre Gaucher.

La vie a ses mystères,
que le destin dévoile.

Remerciements

Avant tout développement, il m'est indispensable de remercier ceux qui m'ont beaucoup appris et accompagner tout au long ce périple.

Je remercie tout d'abord Christine Marcatté et Gwenael Le Lay pour m'avoir accueillie au sein du laboratoire OPERA d'Orange Labs et pour m'avoir intégrée à leur équipe.

Merci à Catherine Colomes de m'avoir recruté pour ce projet et pour m'avoir appris à ne pas rompre face à l'adversité.

Je remercie chaleureusement Mathieu Paquier, mon encadrant universitaire, qui m'a formée et accompagnée tout au long de cette route parfois rocailleuse avec beaucoup de patience et de pédagogie. Je le remercie infiniment d'avoir su me donner confiance, de m'avoir encouragée et d'y avoir cru lorsque je n'y croyais plus, pour finir sans regret cette aventure.

Un grand merci à Laetitia Gros, Rozenn Nicol, Marc Emerit et Catherine Quinquis pour m'avoir éclairée de leurs lumières lors de périodes nuageuses.

Je remercie sincèrement Bernard Le Tertre qui a mené les tests audio parfaitement ainsi que tous les testeurs qui, de par leurs participations et leurs oreilles aguerries, ont contribué à ces travaux de recherche.

Je remercie sincèrement Samuel Moulin sans qui le doute aurait sûrement primé sur la persévérance. Nos discussions autant scientifiques qu'amicales ont fait naître une belle amitié. Je ne le remercierais jamais assez pour son soutien et son aide précieuse dans un moment décisif pour cette thèse.

Une pensée à l'intention de mes collègues : Julien Libouban, Julie Lassalle, Julien Capobianco, Yves Zango, Maty N'Daye, Jerôme Briard, Quentin Vourch, Marion Peres.

La "breizh team" : Sam, Alain, Vincent, Ju, Elo, Pierre, Micka, Cécilia. Je me souviendrais de nos virés à St Malo, nos descentes de ski, nos moments détente à CP, nos sauts dans le vide, nos apéros pétanques, nos soirées poker/jeux... Merci pour votre amitié!!!

Mes chers parents, je m'excuse du fond du coeur pour tous les moments où j'ai reporté mes doutes et mes angoisses sur vous. Merci pour votre écoute et votre soutien indéfectible.

Toi, mon ange, qui m'a encouragée et soutenue pour conclure ce marathon interminable, Merci.

Résumé

Aujourd'hui, les technologies de captation et de restitution sonore se développent dans le but de diffuser des scènes avec un rendu spatialisé. Avant leur diffusion, les extraits sonores peuvent être évalués en terme de qualité par des méthodes recommandées par l'Union Internationale des Télécommunications (évaluation des codecs de compression, procédés de prise ou restitution sonore...). Cependant, ces standards d'évaluation montrent certaines faiblesses notamment en ce qui concerne les attributs de qualité à évaluer. La dimension spatiale n'est pas prise en compte spécifiquement. Dans ce travail, une méthodologie dédiée à l'évaluation de la qualité de l'audio spatialisé est mise en place notamment pour répondre aux biais identifiés. De par l'utilisation d'une catégorisation libre et d'une analyse multidimensionnelle, vingt huit attributs ont été catégorisés en trois familles d'attributs : le *Timbre*, l'*Espace* et les *Défauts*. Ces trois attributs généraux ont été inclus dans un test d'écoute. Celui-ci se déroule en deux phases : l'évaluation de la qualité globale suivie de l'évaluation des trois attributs simultanément sur une même interface. Les tests sont réalisés sans référence explicite, le fichier original constitue une référence cachée. De plus, trois signaux audio, dit ancrages, spécifiques à chacun des trois attributs ont été définis puis superposés pour définir un ancrage unique triplement dégradé. La méthode a été testée à la fois sur un système de restitution au casque avec des contenus binauraux mais également sur un système multicanal 5.1. L'évaluation de stimuli de qualité intermédiaire est préconisée ainsi que des contenus présentant un effet spatial prononcé. L'évaluation multicritère a montré son intérêt dans certaines conditions et permet ainsi d'identifier les caractéristiques qui sont dégradées. Les attributs *Défauts* et *Timbre* ont montré un poids influant sur la qualité globale tandis que le poids de l'attribut *Espace* est plus discutable.

Mots-clefs

Évaluation subjective, qualité, perception sonore, spatialisation, multicanal 5.1, binaural, ancre, MUSHRA, évaluation multicritère, attribut perceptif.

Abstract

Method for the subjective evaluation of spatial sound quality

Nowadays, recording and restitution technologies focus on a spatial rendering of sound. Before their broadcast, the quality evaluation of sound excerpts is often necessary. Methods recommended by the international telecommunication union denote some weaknesses about sound attributes to be evaluated. For example, spatial dimension is barely taken into account. A methodology dedicated to the assessment of spatial audio quality is proposed in order to avoid some biases. With a free categorization and a multidimensional scaling, 28 attributes were clustered in three families : *Timbre*, *Space* and *Defects*. These three categories were included in a listening test split into two sessions : first, the assessment of overall quality and then, the evaluation of the three categories presented simultaneously on a same interface. Tests were conducted without explicit reference, but, the original version was considered as a hidden reference. Moreover, three specific anchors, each one associated to dedicated categories, were defined and then were mixed to define a unique anchor impaired in three ways. The method was tested on a 5.1 system and on binaural contents with headphone restitution. Intermediate quality of contents is recommended as well as contents with relevant spatial effects. The interest of a multicriteria assessment is to identify which properties of sound are impaired. Linear regression shows that *Defects* and *Timbre* attributes have influential weight on overall quality while the weight of *Space* attribute is more dubious.

Keywords

Subjective evaluation, quality, sound perception, spatialization, multichannel 5.1, binaural, anchor, MUSHRA, multicriteria evaluation, perceptive attributes

Table des matières

Introduction générale

Aujourd'hui, les technologies tendent à nous plonger dans le réel de manière virtuelle. Cet oxymore est le coeur des recherches réalisées de nos jours. Nous voyons en 3D, nous entendons en 3D. Les technologies actuelles essaient de reproduire cette perception en développant la spatialisation sonore et elles s'installent petit à petit dans notre quotidien.

La reproduction sonore spatialisée se développe dans de nombreux contextes, que ce soit pour la musique, le cinéma ou encore la visioconférence... Les systèmes de spatialisation sonore sont multiples. Le son multicanal, par exemple, est présent dans de nombreux foyers avec le système d'écoute 5.1. D'autres technologies de spatialisation moins répandues existent : l'holophonie, l'ambisonic, ... Une de ces techniques, la technologie binaurale, permet une écoute spatialisée au casque. Cette dernière ne nécessite que deux canaux comme pour la stéréophonie, cependant, l'impression de spatialisation est nettement accrue.

La nécessité d'avoir des formats audio adaptés à ces technologies est devenue une priorité. Avant leur diffusion et afin de proposer des services de meilleure qualité (choix d'un codage adapté à un service de diffusion), ces formats peuvent faire l'objet d'évaluation de qualité. Deux approches existent. La première regroupe les méthodes objectives fondées sur des mesures physiques des signaux. La seconde est l'évaluation subjective, qui se base sur la perception des auditeurs évaluée lors de tests d'écoute.

L'Union Internationale des Télécommunications (UIT) est un institut de normalisation qui recommande, entre autres, des méthodes pour évaluer subjectivement la qualité sonore dans le contexte de séquences sonores fortement ou faiblement dégradées par des systèmes de codages audio. Les deux méthodes les plus utilisées sont la recommandation ITU-R BS.1116 (1997) et la méthode appelée MUSHRA (ITU-R BS.1534, 2003). Ces standards, reconnus par la communauté scientifique, sont les principaux mis en place lors de la conception de tests d'écoute. Cependant, des études ont révélé que ces méthodes comportaient certains biais : l'échelle de notation, les attributs à évaluer, les ancrages... De plus, la dimension spatiale du son n'est pas prise en considération spécifiquement.

Le but des travaux présentés dans ce manuscrit est de mettre en place une méthode pour évaluer la qualité sonore adaptée aux sons spatialisés afin de réduire les biais identifiés. Dans un premier temps, les axes perceptifs autres que la qualité globale sont définis. Ensuite, ces axes sont intégrés dans un test d'écoute et leur pertinence est évaluée sur différents systèmes de restitution spatialisée.

Ce mémoire s'articule autour de huit chapitres. Le premier présente les deux technologies de restitution sonore spatialisée sonore utilisées lors des tests d'écoute : le système 5.1 et la technologie binaurale. Les principes des différentes méthodes d'évaluation existantes

et leurs limitations y sont également détaillés.

Dans le second chapitre, les catégorisations et l'élicitation d'attributs pour qualifier le son sont explicitées et deux d'entre elles sont utilisées pour catégoriser une liste d'attributs afin de définir des axes perceptifs de la qualité dans le but de les inclure dans un test d'écoute.

L'évaluation de la qualité sur un système 5.1 selon les trois catégories mises en avant fait l'objet du troisième chapitre. Le mode de présentation de ces attributs aux auditeurs, successivement ou simultanément, est également étudié.

La méthode est ensuite appliquée à une restitution au casque avec des contenus binauraux dans le chapitre quatre. Trois ancrages spécifiques à chaque attribut sont inclus dans le test. Il s'avère que l'ancrage spatial choisi n'a pas joué son rôle.

Le chapitre cinq propose donc un test pour définir un ancrage spatial adapté à la méthode de test.

L'ancrage spatial défini précédemment fait l'objet d'une validation dans le chapitre six.

Un nombre élevé de stimuli dans un test peut présenter un biais. Un ancrage unique triplement dégradé est étudié dans le chapitre sept. Il remplace les trois ancrages spécifiques à chaque axe perceptif.

Enfin, le chapitre huit décrit la méthodologie de tests subjectifs pour évaluer les sons spatialisés dans le contexte des systèmes de codage.

Chapitre I

L'évaluation de la qualité en audio spatialisé

Dans le contexte des systèmes de codages audio (compression de débit avec pertes de qualité), la qualité des contenus nécessite d'être évaluée. Ce chapitre présente les principales méthodes utilisées pour évaluer la qualité audio ainsi que les biais induits par l'emploi de ces méthodes. Tout d'abord, la perception de l'espace et les systèmes de restitution sonore spatialisée multicanaux et binauraux sont décrits.

I.1 La localisation auditive

La localisation auditive (azimut, élévation, distance) est permise grâce au traitement de différents indices : indices interauraux, indices monauraux, rapport champs direct/champs diffus.

I.1.1 Les indices interauraux

La théorie *Duplex* proposée par Rayleigh (1907) permet de décrire de manière simple les mécanismes de perception en azimut (plan horizontal) par l'utilisation de deux indices perceptifs interauraux : l'ITD (Interaural Time Difference) et l'ILD (Interaural Level Difference).

L'ITD caractérise la différence interaurale de temps d'arrivée d'une onde acoustique provenant d'une source à une position donnée. Le retard occasionné est la conséquence de la différence entre le trajet "source - oreille ipsilatérale" (OA) et le trajet "source - oreille controlatérale" (OB), représentés sur la figure I.1 (Moulin, 2011).

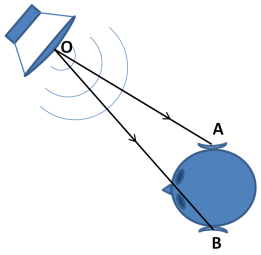


Fig. I.1 – Représentation des signaux incidents aux oreilles ipsilatérale (point A) et controlatérale (point B) (Moulin, 2011).

L'ITD porte essentiellement l'information du degré de latéralisation de la source sonore et ce jusqu'à une fréquence d'environ 1500 Hz (cette fréquence limite est dépendante de la distance interaurale).

L'ILD caractérise les différences de niveau sonore perçues entre chaque oreille pour une source à une position donnée. Il apparaît comme un indice potentiel de latéralisation, complémentaire de l'ITD. Cet indice permet une perception du degré de latéralisation pour les fréquences où l'ITD n'est plus efficace (pour $f \geq 1500Hz$).

La théorie *Duplex* proposée par Lord Rayleigh permet de comprendre les mécanismes de la localisation en azimut mais l'utilisation exclusive de l'ITD et de l'ILD ne permet pas au système auditif une perception de l'élévation.

I.1.2 Les indices monauraux

Les IS (indices spectraux) sont des indices monauraux. Ils prennent en compte les effets des réflexions et diffractions des ondes sonores causés par leurs interactions avec le corps d'un auditeur. La morphologie d'un individu engendre une modification spectrale du signal source. L'auditeur interprète ce filtrage anatomique à chaque situation donnée et le compare avec les filtrages stockés en mémoire. Il peut ensuite en déduire la position de la source par reconnaissance de forme. Les IS permettent ainsi une localisation en élévation et participent à diminuer la confusion avant/arrière.

I.1.3 Les HRTF

Les HRTF (Head Related Transfer Function) sont les fonctions de transfert qui décrivent la propagation acoustique entre la source sonore et les oreilles de l'auditeur. Ces HRTF rassemblent sous une forme compacte l'ensemble des indices monauraux et binauraux mis à disposition du système auditif pour localiser les sons. Elles constituent le concept fondamental des technologies binaurales. La position d'une source sonore est encodée par la fonction de transfert associée à sa direction, et qui traduit l'ensemble des phénomènes de propagation des ondes acoustiques entre la source et l'entrée des conduits auditifs (Moulin, 2011), à savoir :

- la propagation en champ libre,
- la diffraction par la tête de l'auditeur (Duda et Martens, 1998), (Algazi *et al.*, 2001),
- les réflexions sur les épaules et le haut du torse de l'auditeur (Algazi *et al.*, 2002a), (Algazi *et al.*, 2002b),
- les résonances liées à la forme du pavillon (Batteau, 1967), (Shaw et Teranishi, 1968), (Hebrank et Wright, 1974).

Les HRTF sont entièrement déterminées par la morphologie d'un individu. D'un individu à l'autre, la fréquence et l'amplitude des pics et des creux des HRTF sont décalées et leur nombre varie. Lorsqu'un sujet est soumis aux HRTF d'un autre individu, sa perception de la localisation de sources sonores est fortement perturbée. Il apparaît alors une augmentation des confusions avant/arrière, une perception intracrânienne, des distorsions de la localisation en élévation et enfin une perte de frontalisation (Hofman *et al.*, 1998).

Les mesures des HRTF sont effectuées dans une chambre anéchoïque. Elles sont réalisées à l'aide de microphones placés dans chaque conduit auditif d'un auditeur. Des haut-parleurs sont placés autour du sujet, pour chacune des positions, un signal est émis et une

mesure est effectuée. La multitudes de positions de la source engendre des temps de mesure extrêmement longs (plusieurs heures pour 1000 HRTF par exemple) ce qui représente une véritable épreuve pour le sujet qui doit rester totalement immobile.

La définition physique des HRTF gauche et droite H_L et H_R est donnée par l'équation (I.1),

$$H_{L,R}(r,\theta,\varphi)(j\omega) = \frac{\phi_{L,R}(r,\theta,\varphi)(j\omega)}{\phi_0(j\omega)}, \tag{I.1}$$

où (r,θ,φ) désigne la position en coordonnées sphériques du haut-parleur de mesure dans le référentiel auditeur, ϕ_L et ϕ_R sont les pressions sonores mesurées à l'entrée des conduits respectivement gauche et droit, et ϕ_0 est la pression acoustique mesurée à la position du centre de la tête, le sujet étant absent (Moulin, 2011).

Les HRTF contiennent tous les indices de localisation de sources sonores d'un individu et permettent le calcul des ITD ou ILD.

Les indices perceptifs interauraux (ITD et ILD) et monauraux (IS) permettent donc la localisation de sources sonores en azimut et en élévation.

I.2 La restitution sonore spatialisée

La stéréophonie peut être considérée comme l'un des premiers systèmes de spatialisation. Elle s'est particulièrement développée à partir des années cinquante. Le terme "Stéréo" vient du grec et signifie ferme, solide qui par extension au sens de "volume" donne une image d'espace à trois dimensions (wiktionary, 2014). La restitution peut être faite au casque ou sur deux haut-parleurs. Son principe repose sur des différences de temps et d'intensité entre les deux canaux. Depuis, dans le but d'amplifier l'effet de spatialisation, les technologies multicanales sur haut-parleurs et la technologie binaurale au casque se sont développées.

I.2.1 Le système 5.1

Le système 5.1, "home cinéma", est un système comprenant cinq haut-parleurs et un caisson de basse. Les notations suivantes sont utilisées (tableau I.1) :

Tab. I.1 – Notation des haut-parleurs.

abréviation	Haut-parleurs (HP)
L	HP avant gauche
R	HP avant droit
C	HP avant centre
Sub	Caisson de basse
Ls	HP arrière gauche
Rs	HP arrière droit

La disposition des haut-parleurs respecte la recommandation ITU-R BS.775-2 (2006) qui positionne le système de la manière suivante (figure I.2) :

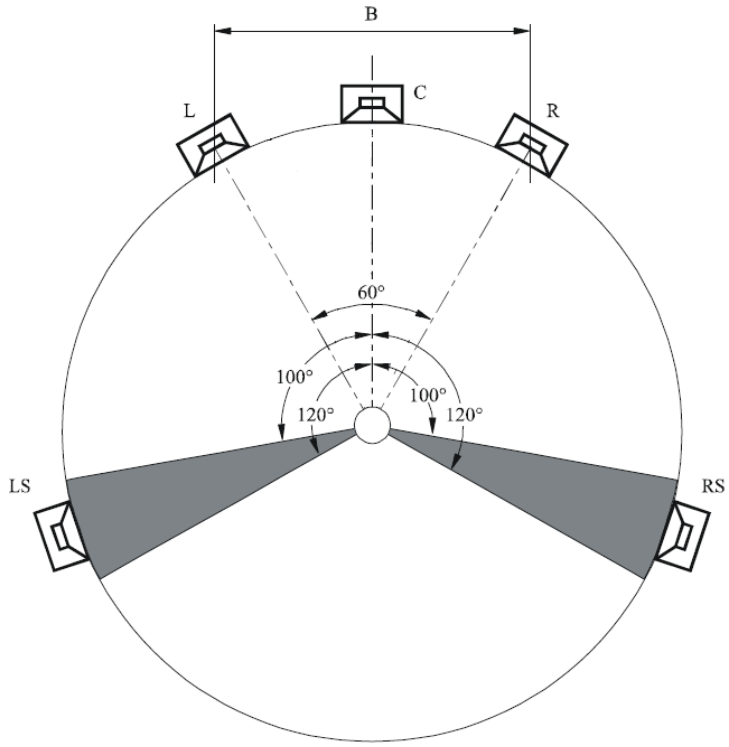


Fig. I.2 – Configuration 5.1 (ITU-R BS.775-2, 2006).

La largeur de base B entre les haut-parleurs L et R est comprise de préférence entre 2 et 3 mètres selon la recommandation ITU-R BS.1116 (1997) et peut atteindre 5 mètres pour des locaux appropriés. La configuration spatiale des enceintes d'un système 5.1 est d'une importance primordiale car elle conditionne directement la qualité d'écoute et le réalisme des effets sonores. Le point d'écoute dit de référence se nomme le "sweet spot" et se situe au centre du cercle sur lequel sont positionnées les cinq enceintes (ITU-R BS.775-2, 2006). Les enceintes doivent être placées à la hauteur des oreilles de l'auditeur.

Le caisson (.1) reproduit le canal Lfe (low-frequency effects) et éventuellement, la partie basse du spectre des cinq autres canaux (bass management). Le canal Lfe permet d'améliorer la restitution des basses fréquences pour des effets spéciaux par exemple. Il est toutefois optionnel (ITU-R BS.775-2, 2006).

I.2.2 La reproduction binaurale

La reproduction binaurale est un procédé qui permet un rendu spatialisé tridimensionnel au casque reposant sur les principes psychophysiologiques de l'audition. La reproduction sonore, réalisée au niveau des conduits auditifs des auditeurs, provoque l'illusion d'être immergé dans une scène sonore en percevant les sources dans un espace tridimensionnel. En ce sens, cette technique se rapproche au plus près de l'écoute naturelle.

Il existe deux formes d'encodage pour la technologie binaurale basées sur les indices de localisation.

– encodage naturel : les signaux binauraux sont enregistrés par une paire de microphones placée à l'entrée des conduits auditifs d'un individu ou d'un mannequin (tête artificielle)

– encodage artificiel : les signaux binauraux sont obtenus par synthèse binaurale en convoluant un signal monophonique représentant le signal émis par la source sonore par une paire de filtres modélisant les HRTF associées aux oreilles gauche et droite en relation avec une position de source donnée.

L'encodage artificiel permet d'obtenir des signaux par synthèse binaurale. Il peut remplacer l'encodage naturel rendu complexe par la mesure des HRTF. En effet, la mesure acoustique des HRTF est coûteuse et laborieuse. De plus, leur caractère individuel constitue une réelle contrainte. Pour l'encodage artificiel, les signaux sont créés de manière synthétique dans le but de donner à l'auditeur l'impression d'une écoute naturelle et de générer le champ acoustique correspondant. La synthèse binaurale consiste à créer une source sonore virtuelle en convoluant le signal source par la paire de HRTF associée à la position à simuler. Des filtres binauraux sont utilisés pour modéliser les HRTF. Le modèle le plus commun se compose d'un filtre à phase minimale qui reproduit le module spectral de la HRTF, et d'un retard pur qui représente l'information temporelle contenue dans les HRTF (Kistler et Wightman, 1992; Kulkarni *et al.*, 1995).

Le développement des technologies binaurales est ralenti par le caractère individuel des HRTF. Des études récentes apportent des solutions partielles au travers de diverses méthodes (Pernaud, 2003; Busson, 2006). Par exemple, Guillon (2009) propose dans ses travaux de thèse une méthode d'interpolation particulière permettant de reconstruire des HRTF à partir d'un nombre réduit de mesures sur l'individu. Une autre solution proposée par Guillon consiste à adapter un jeu de HRTF, issu d'une base de données, par le biais d'une comparaison morphologique des auditeurs. Des modèles de calcul de filtres binauraux individualisés à l'auditeur, à la fois en termes d'ITD et d'IS, ont été proposés dans les travaux de Busson (2006). La finalité de ces travaux vise l'intégration d'un moteur de spatialisation binaurale dans des applications grand public dans le contexte étendu des télécommunications d'aujourd'hui.

I.3 Les méthodes d'évaluation subjective de la qualité sonore

Pour satisfaire un niveau d'exigence, il est indispensable que les contenus audio soient, avant toute diffusion, évalués en terme de qualité. Dans le milieu industriel, il est nécessaire de se référer à des méthodes d'évaluation normalisées pour donner de la valeur aux résultats obtenus et aux choix qui en découlent. En effet, en suivant une méthodologie recommandée et bien détaillée, les conditions et procédures d'évaluation sont identiques à tout expérimentateur. Les tests sont alors reproductibles et peuvent être comparés et discutés. Des organismes comme l'Union Internationale des Télécommunication (UIT) ou L'Union Européenne de Radio-télévision (UER) définissent des normes à suivre notamment pour l'évaluation de qualité. Ces normes sont appliquées à l'évaluation de contenus sonores qui ont subi des dégradations notamment par l'application de codages audio.

Une procédure généralisée peut être établie pour définir une méthodologie de test. Lawless et Heymann (1998) ont décrit une procédure d'évaluation en trois temps (figure I.3) :

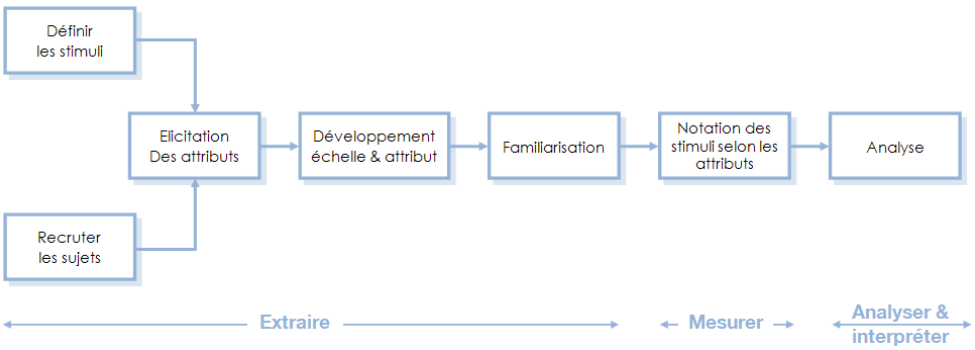


Fig. I.3 – Principales étapes d'une méthode de test (Pedersen et Zacharov, 2008).

* Extraire : cette première étape regroupe la composition du panel, du corpus de stimuli, les paramètres à évaluer et les échelles
* Mesurer : cette seconde partie correspond à l'évaluation des stimuli
* Analyser et interpréter : cette troisième et dernière phase s'intéresse aux choix de la méthode d'analyse (statistique ou linguistique) et à l'interprétation des résultats.

Dans la recommandation EBU Tech 3286 (1997), il est écrit qu'une méthode d'évaluation subjective de la qualité audio doit répondre à cinq prérequis pour être pertinente :
* la composition du panel d'écoute
* les conditions d'écoute
* les paramètres à évaluer
* l'échelle de notation
* la méthode de récolte et d'analyse des résultats.

Dans le but de définir un standard d'évaluation, il est nécessaire de répondre précisément à ces critères pour faciliter l'élaboration de tests d'écoute subjectifs.

I.3.1 ITU-R BS.1534 (MUSHRA)

La méthode appelée communément MUSHRA (MUltiple Stimilus with Hidden Reference and Anchors) est définie par la recommandation ITU-R BS.1534 (2003). Il existe des applications qui, encore aujourd'hui, ne peuvent diffuser en haute qualité comme par exemple la diffusion par Internet. Cette méthode est dédiée à l'évaluation subjective des systèmes de codage de niveaux de qualités intermédiaires. Il a été montré que cette méthode conduit à des résultats fiables et convergeant sur peu d'auditeurs experts (Soulodre et Lavoie, 1999; EBU, 2000b,a). La recommandation est applicable sur tout dispositif de reproduction, système monophonique, stéréophonique ou multivoie au casque ou sur haut-parleurs.

Le panel d'auditeurs
Bien que la méthode ne soit appliquée qu'à des niveaux de qualité intermédiaire, la norme recommande de faire appel à des sujets expérimentés. Ces personnes ont l'habitude de ce type de test, d'écouter le son de manière critique et possèdent des capacités auditives normales au sens de la norme ISO.389 (1985). La méthode préconise une vingtaine de participants pour valider les résultats.

L'échelle
L'échelle de notation est une échelle continue de qualité allant de 0 à 100 avec 5 intervalles nommés mauvais [0-20], médiocre [20-40], assez bon [40-60], bon [60-80] et excellent [80-100]. Cette échelle provient de la recommandation ITU-R BT.500-11 (2002) utilisée pour l'évaluation de la qualité d'image.

Les stimuli
Les extraits sélectionnés pour un test doivent être "critiques" pour permettre de différencier les systèmes testés. Les extraits choisis ne doivent pas excéder vingt secondes pour limiter la longueur d'un test d'écoute, la fatigue engendrée pour les auditeurs et pour diminuer l'effet de mémoire à court terme. Lors d'une session, il est recommandé de proposer au maximum quinze versions différentes d'un extrait avec au minimum cinq extraits différents. Parmi ces quinze items, il faut compter le signal original (la référence explicite), la référence cachée et au moins un signal d'ancrage. Cet ancrage ou repère caché est le résultat du filtrage passe bas du signal non dégradé (la référence) coupé à $3.5kHz$. D'autres ancrages peuvent être utilisés comme par exemple la limitation de la largeur de bande à 7 ou 10kHz, l'image stéréo réduite, du bruit supplémentaire, des pertes de paquets, des pertes de signal, etc... Par exemple, si un test contient 5 codages audio à évaluer, un minimum de 8 signaux différents seront présentés lors de la phase de notation incluant le signal de référence, les 5 signaux dégradés par les codages, 1 signal de référence cachée et 1 signal d'ancrage caché.

Comme énoncé précédemment, les versions ou objets évalués, doivent avoir subi des dégradations moyennes ou fortes (Soulodre et Lavoie, 1999). Les dégradations sont marquées et la détection des altérations n'est pas difficile. En revanche, la recommandation ITU-R BS.1116 (1997), détaillée dans la section I.2.2, est dédiée à l'évaluation de la qualité audio pour des dégradations faibles c'est-à-dire pour des systèmes haute qualité. Il est spécifié que la qualité des versions évaluées avec la méthode MUSHRA doit figurer dans la moitié inférieure de l'échelle proposée par la norme ITU-R BS.1116 (1997).

Les attributs évalués
Quel que soit le format de restitution évalué (mono, stéréo, multicanal), il est préconisé d'évaluer, pour chaque test, la *qualité audio de base* qui est par définition "la caractéristique unique et globale pour évaluer toutes les différences décelées entre la référence et l'objet du test". La recommandation propose d'autres caractéristiques à évaluer pour les systèmes autre que monophoniques. Pour les systèmes stéréophoniques, il s'agit de la *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". Pour les systèmes multicanaux, les attributs et leurs définitions sont :

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

Il est précisément indiqué que, pour une évaluation multicritère, les attributs doivent être évalués lors de séances différentes. En effet, les auditeurs peuvent éprouver des diffi-

cultés à évaluer plusieurs caractéristiques en même temps et les résultats peuvent perdre en fiabilité. Il n'y a pas d'informations sur la provenance de ces attributs ce qui peut expliquer le fait qu'ils soient rarement utilisés.

La procédure de test
MUSHRA est un test de comparaison multiple. La figure I.4 présente une interface de test possible pour cette méthode.

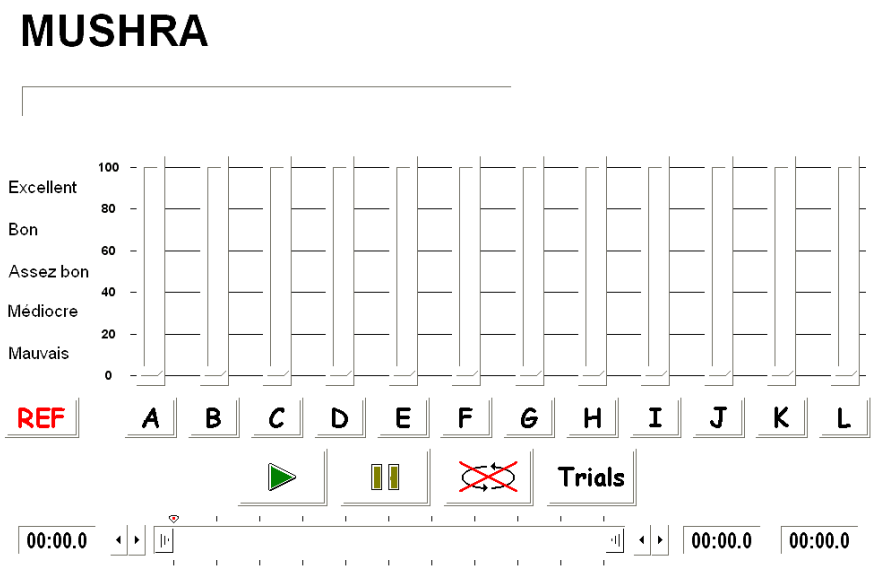


Fig. I.4 – Exemple d'interface issue de CRC (SEAQ) pour la méthode ITU-R BS.1534.

Lors de l'évaluation d'un extrait, la référence explicite est notée REF et toutes les dégradations, y compris l'ancre et la référence cachée, sont présentées simultanément aux auditeurs et sont réparties aléatoirement derrière les lettres A, B, C, D... Chaque auditeur peut écouter les extraits sonores autant de fois que nécessaire dans l'ordre de son choix. Il a la possibilité de commuter d'une version à une autre et donc de procéder directement à toutes les comparaisons. En utilisant l'échelle de qualité décrite précédemment, l'auditeur attribue une note de qualité perçue en plaçant le curseur sur l'échelle de notation. Étant donné la présence d'une référence cachée, au moins une version doit être notée à 100, la note maximale de l'échelle. Une fois satisfait des notes attribuées à chaque stimulus, l'auditeur peut accéder à l'extrait suivant.

La postsélection des sujets
Les résultats d'un sujet peuvent être source de biais. Une procédure de rejet est mise en place pour mettre de coté les sujets trop ou pas assez critiques, se détachant trop du groupe. Le but est de pouvoir éliminer des "observations aberrantes". Deux méthodes sont proposées pour procéder à une postsélection des sujets. La première porte sur la fiabilité et la constance du sujet dans sa notation. La seconde se base sur l'écart entre l'évaluation individuelle et la moyenne des évaluations.

L'analyse des résultats
Les analyses statistiques reposent sur le calcul des moyennes et des intervalles de confiance à 95% du panel d'auditeurs après postsélection.

I.3.2 ITU-R BS.1116

La recommandation ITU-R BS.1116 (1997) est une méthodologie de test utilisée pour l'évaluation de la qualité de contenus audio présentant de faibles dégradations. La méthode BS.1116 et la méthode MUSHRA ont plusieurs points communs (les conditions d'écoute, les attributs, les dispositifs de reproduction, la postsélection des sujets), cependant l'échelle de notation et la procédure de test sont totalement différentes.

Le panel d'auditeurs
Les sujets recrutés sont au nombre de 20 environ et doivent être capables de déceler des dégradations infimes dans des stimuli audio et les évaluer. La présélection des participants peut se faire sur la base de tests audiométriques ou de l'expertise acquise lors de tests d'écoute réalisés précédemment.

L'échelle
L'échelle de notation est une échelle continue de dégradation avec 5 échelons : 1 (dégradation très gênante), 2 (dégradation gênante), 3 (dégradation légèrement gênante), 4 (dégradation perceptible mais non gênante) et 5 (dégradation imperceptible). La résolution est d'une décimale.

Les stimuli
La durée des séquences est de 10 secondes à 25 secondes. La version originale est présente à chaque évaluation. Les versions évaluées doivent présenter des dégradations faibles.

La procédure de test
La figure I.5 illustre un exemple d'interface utilisée pour cette méthode. Un auditeur doit évaluer trois items notés A, B et C. A est la référence explicite. B et C sont aléatoirement la référence cachée ou la version dégradée. Les extraits peuvent être écoutés autant de fois que nécessaire. Pour chaque essai, l'auditeur identifie la référence cachée notée à 5 par défaut. Ensuite il évalue la version testée en utilisant l'échelle de dégradation. Concernant l'analyse statistique, le modèle ANOVA est souvent appliqué.

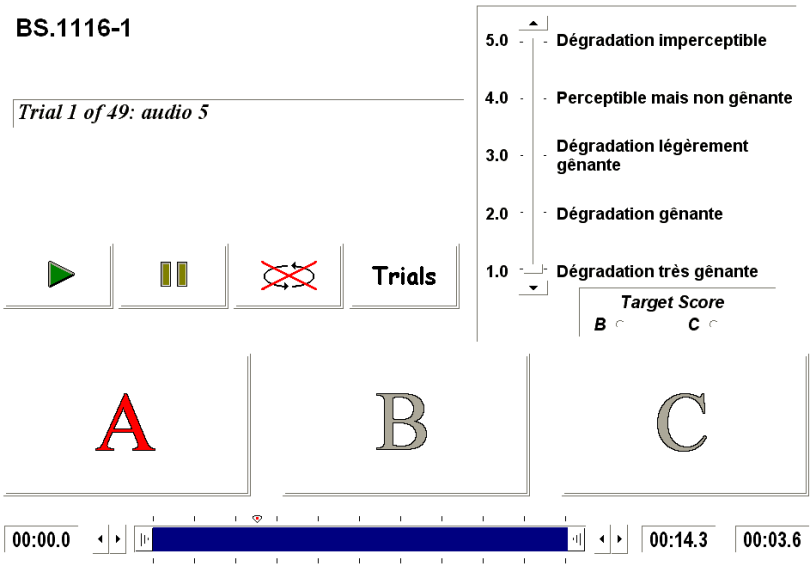


Fig. I.5 – Exemple d'interface de CRC (SEAQ) pour la méthode ITU-R BS.1116 (1997).

La méthode d'évaluation ITU-R BS.1116 (1997) décrit les dispositifs de reproduction comprenant les hauteurs, positions et orientations des hauts parleurs. Les conditions d'écoute à satisfaire comme les propriétés acoustiques et géométriques du local d'écoute, le niveau sonore, y sont également stipulées.

La précision, tant au niveau de la procédure de test que des conditions d'écoute, permet une totale reproductibilité des tests réalisés en suivant cette méthode.

I.3.3 EBU Tech 3286

En 1997, l'Union Européenne des Radio-télévision (UER) propose la méthode EBU Tech 3286 (1997). Celle-ci est dédiée à l'évaluation de contenus de musique classique, par exemple l'opéra, la musique de chambre ou encore la musique symphonique... Cependant elle peut être appliquée à d'autres contenus musicaux tant qu'ils consistent en une performance acoustique en live réalisée dans un espace réel. Cette méthode ne s'applique pas à la musique électrique/électronique, à la parole ou à la production cinématographique.

Le panel d'auditeurs
Le panel est composé d'auditeurs experts qui ont été entraînés à utiliser cette méthode d'évaluation de qualité. Il est recommandé que les participants aux tests travaillent dans le domaine de la production sonore ou aient une expérience approfondie dans l'écoute des sons de manière professionnelle. Les participants doivent avoir une audition normale au sens de la norme ISO.389 (1985) et une maîtrise du langage utilisé lors des évaluations.

Les conditions d'écoute
Les conditions d'écoute à respecter dans cette méthode sont décrites avec détails dans la norme EBU Tech 3276 (1997). Par exemple, sont explicités les configurations des systèmes de diffusion, les mesures acoustiques comme le temps de réverbération, la réponse en fréquence de la salle, le niveau d'écoute...

Les attributs de qualité
Le protocole de test prend en compte plusieurs attributs de qualité dans l'évaluation. Six catégories d'attributs sont incluses pour la stéréophonie et huit pour la restitution multicanale (EBU Tech 3286 Supp. 1, 2000). Voici la liste des huit attributs proposés et leurs définitions :

- Qualité de l'image frontale : les images frontales donnent l'impression d'avoir une distribution directionnelle correcte et appropriée.
- Qualité sonore des côtés et arrières : les arrières et les cotés donnent l'impression d'être équilibrés correctement et de manière appropriée.
- Impression spatiale : la prestation donne l'impression de se dérouler dans un environnement approprié.
- Transparence : tous les détails de la prestation sont clairement perçus.
- Balance sonore : les sources individuelles et les ambiances apparaissent correctement balancées dans l'ensemble sonore.
- Couleur sonore : représentation précise des diverses caractéristiques sonores de(s) (la) source(s).
- Absence de bruit et de distorsion : présence des diverses nuisances (bruits électriques et acoustiques, bruits de public, erreurs numériques, distorsion...).
- Impression générale : une moyenne pondérée des paramètres proposés qui prend en

compte l'intégrité de l'image sonore globale et les interactions entre les différents paramètres.

Pour chaque catégorie, plusieurs attributs sont proposés comme par exemple pour l'impression spatiale : la réverbération, l'enveloppement, la profondeur, la réalité spatiale, la balance acoustique, la taille apparente de la salle...

L'échelle
L'échelle d'évaluation est une échelle catégorielle avec six échelons (tableau I.2). Le choix d'une échelle en six points a été fait dans le but de forcer un rang positif ou négatif, soit bon, soit mauvais, il n'y pas de rang neutre.

Tab. I.2 – Échelle catégorielle de la méthode EBU Tech 3286.

Échelon	Qualité	Impression
1	Mauvaise	Défauts très gênants
2	Insuffisante	Trop de défauts gênants
3	Suffisante	Plusieurs défauts gênants
4	Bonne	Quelques défauts mineurs
5	Très bonne	Quelques défauts audibles non gênants
6	Excellente	Pas de défaut audible

La procédure de test
Les auditeurs écoutent les extraits enregistrés en amont sur cassettes DAT selon la recommandations R64. Il n'y a pas de son de référence. Une pause de deux minutes maximum est réalisée entre chaque item. Ce temps leur permet de réaliser leur évaluation. La figure I.6 donne un exemple de fiche d'évaluation donnée aux sujets. Après chaque extrait, ils cochent le grade de qualité perçue et notent leurs commentaires pour chacun des huit attributs.

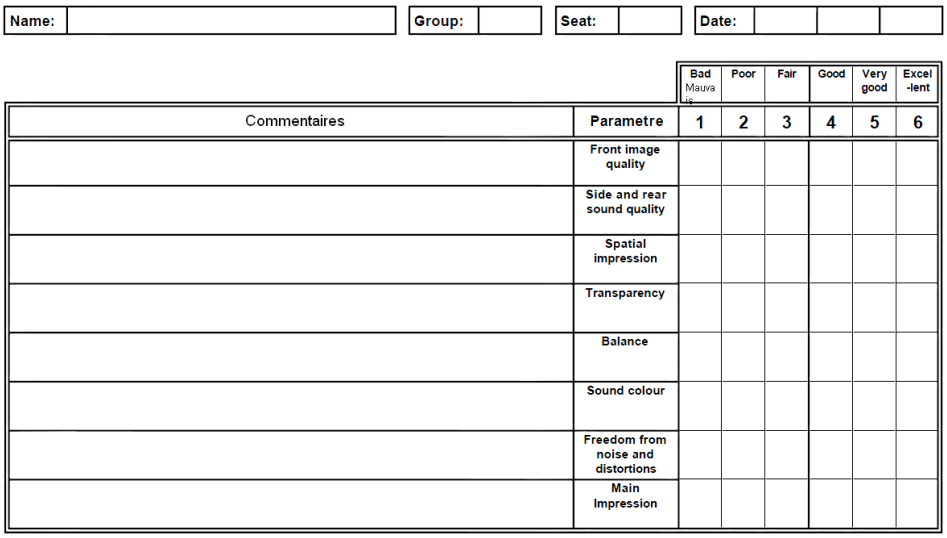


Fig. I.6 – Exemple de fiche d'évaluation donnée aux auditeurs (EBU Tech 3286, 1997).

Les résultats
Les résultats sont analysés avec des méthodes non-paramétriques parce que l'échelle est discrète. Leur représentation se fait sous la forme d'un diagramme radar (figure I.7).

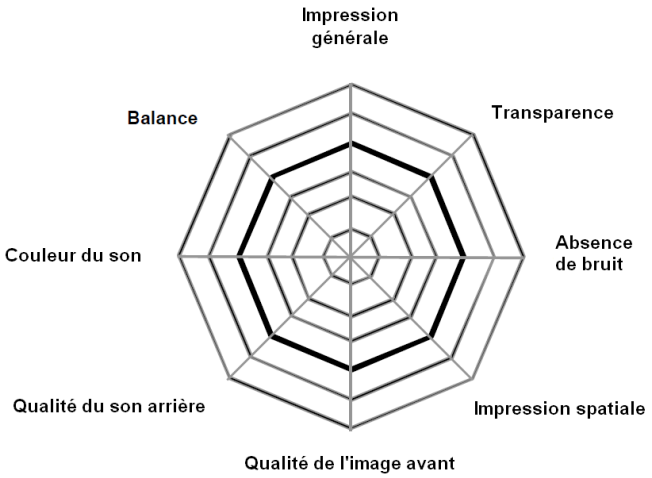


Fig. I.7 – Représentation des résultats sous la forme d'un diagramme radar (EBU Tech 3286, 1997).

I.3.4 ITU-R BS.1284

La recommandation ITU-R BS.1284 (2003) regroupe les différentes méthodes d'évaluation subjective de la qualité telles que la comparaison par paire et la comparaison multiple. Trois échelles continues sont proposées (figure I.8).

Qualité	Dégradation	Comparaison
5 Excellent	5 Imperceptible	3 Nettement supérieur
4 Bon	4 Perceptible, mais non gênante	2 Supérieur
3 Assez bon	3 Légerement gênante	1 Légèrement supérieur
2 Médiocre	2 Gênante	0 Comparable
1 Mauvais	1 Très gênante	-1 Légèrement plus mauvais
		-2 Plus mauvais
		-3 Nettement plus mauvais

Fig. I.8 – Echelles proposées par la recommandation ITU-R BS.1284.

C'est une méthode "à la carte" qui propose des tests sous la forme d'une présentation unique, d'une comparaison par paires ou d'une comparaison multiple chacune incluant ou non une référence avec le choix d'une des trois échelles représentées sur la figure I.8. Les sujets doivent être experts. Cette méthode propose les mêmes attributs que la méthode ITU-R BS.1534 (2003) cependant elle inclut également la liste des critères d'évaluation de qualité de la méthode EBU Tech 3286 (1997).

I.4 Les biais des standards d'évaluation

L'utilisation de méthodologies normalisées pour évaluer les qualités des sons est indispensable pour l'échange, la compatibilité et la comparaison des résultats obtenus. Cependant, des études récentes ont montré que les standards d'évaluation de qualité comportaient certains biais (Zielinski *et al.*, 2007b, 2008).

I.4.1 La qualité audio de base

La qualité audio de base (BAQ) est bien souvent l'unique attribut évalué par les auditeurs lors de tests audio. Les recommandations proposent des attributs supplémentaires mais sans référence sur l'origine et le choix de ces attributs. Ceci peut expliquer le fait qu'ils ne soient jamais utilisés. De plus, avec le développement des technologies de restitution sonore spatialisée, la BAQ semble insuffisante pour évaluer la qualité sonore.

I.4.2 Les ancrages

Les ancrages servent de référence moyenne et/ou basse qualité lors de tests d'écoute. Il existe peu d'informations sur leurs constructions et sur leur pertinence dans un test. Le principal ancrage proposé est le signal filtré à 3.5 kHz. L'ajout d'attributs dans un test d'écoute pourrait engendrer la conception d'ancrages spécifiques à ces attributs.

I.4.3 La référence

Les normes incluent toujours une référence explicite. Les objets à évaluer sont comparés directement avec un fichier dit de référence. Il est considéré comme l'élément de meilleure qualité dans le test d'écoute. Cependant, un objet pourrait être de meilleure qualité que cette référence et l'auditeur ne pourrait pas le notifier. De plus, la présence d'une référence explicite implique une tâche d'évaluation de fidélité à cette référence plutôt qu'une évaluation de qualité entre les objets. Il parait intéressant de réaliser des tests sans référence explicite, en gardant la consigne de noter un objet au plus haut de l'échelle de notation. Cette consigne permet d'utiliser la dynamique haute de l'échelle et de réduire l'effet du choix des extraits sur les versions évaluées.

I.4.4 L'échelle de notation

Il a été montré que les échelles utilisées pour la notation dans les tests d'écoute comportent certains biais (Zielinski *et al.*, 2007a). Deux problèmes majeurs ont été identifiés. Le premier correspond à la non équidistance entre les étiquettes qui constituent l'échelle. Pour rappel, l'échelle de qualité utilisée dans MUSHRA est divisée en cinq intervalles égaux. Jones et McManus (1986) ont montré que les intervalles sémantiques entre les étiquettes ne sont pas équivalents entre eux. Dans leur étude, il était demandé aux sujets de noter graphiquement sur une échelle ayant pour extrémités "Pire imaginable" et "Meilleur imaginable" des adjectifs, dont ceux utilisés par les échelles de notation. Les résultats ont montré que les intervalles entre les termes ne sont pas équidistants. Il faut préciser que cette étude a été réalisée en anglais US. Les échelles normalisées sont utilisées de manière internationale et donc traduites dans plusieurs langues. De ce fait, Jones et McManus ont réitéré leur expérience cette fois-ci en italien, et ont confirmé que l'étalement des adjectifs n'est pas linéaire. Ce même résultat a été vérifié dans d'autres langues comme le français

(ITU-R BT.1082-1, 1990), l'anglais britannique (Watson, 1999), le suédois et le néerlandais (Virtanen *et al.*, 1995; Teunissen, 1996). La figure I.9 résume les résultats obtenus et montre les écarts entre les adjectifs de qualité dans différentes langues.

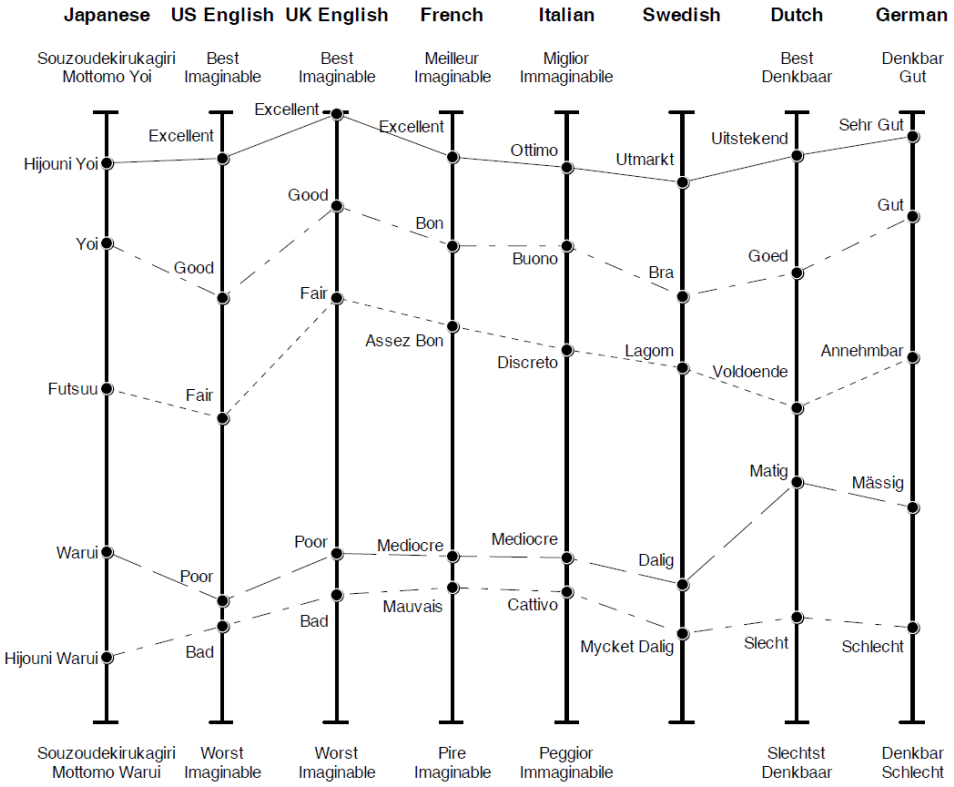


Fig. I.9 – Synthèse des résultats obtenus pour vérifier la linéarité des termes utilisés dans les échelles de qualité (Zielinski *et al.*, 2008).

Les termes "mauvais" et "médiocre" et leurs traductions sont considérés comme très proches dans la partie basse de l'échelle et vérifient un écart important avec le qualificatif "assez bon" situé dans la partie supérieure de l'échelle. De plus, il y a des différences entre les termes considérés comme équivalents d'une langue à l'autre. Ainsi le terme "excellent" en anglais britannique est différent du terme "excellent" en anglais américain. Ils sont linguistiquement identiques mais ils ne valent pas exactement le même degré de qualité. Il en est de même pour chaque étiquette traduite dans différents langages, le terme "assez bon" et ses diverses traductions sont placés à des niveaux différents sur l'échelle. A l'inverse des autres langues, l'allemand (ITU-R BT.1082-1, 1990) et le japonais (Narita, 1993) font apparaître une dispersion uniforme des termes sur l'échelle et valident donc les échelles normalisées.

Au vu des études réalisées, bien que les résultats soient dépendants du pays ou de la langue dans lesquels étaient conduits les tests, l'échelle de qualité est considérée comme non linéaire (Zielinski *et al.*, 2007a).

Le second biais identifié correspond à la notation des sujets influencés par l'emplacement des étiquettes des échelles. En effet, ils ont tendance à noter les stimuli en face de celles-ci alors que l'échelle est continue comme le montre la figure I.10.

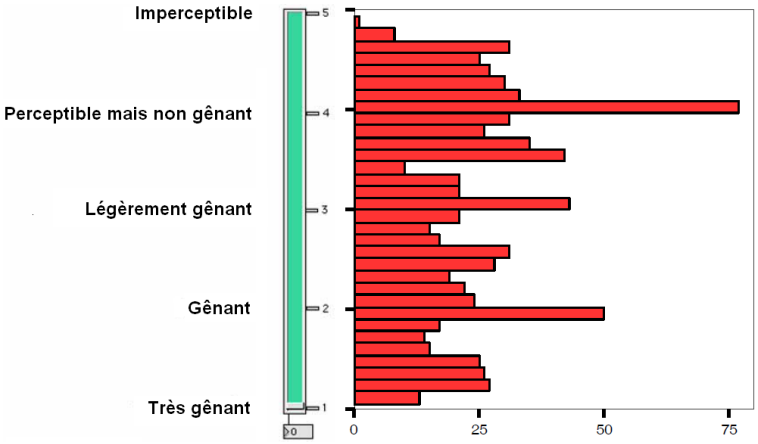


Fig. I.10 – Exemple de dispersion des sujets par l'utilisation de l'échelle de dégradation (Zielinski *et al.*, 2007a).

Du fait de ces biais, Watson (1999) s'interroge sur la popularité des échelles standardisées notamment dans les firmes de télécommunication et argumente en faveur d'une échelle non étiquetée comme alternative. ITU-R BT.1082-1 et ITU-T P.910 (1999) suggèrent une échelle comprenant des labels aux extrémités mais sans dénomination intermédiaire, échelle qui est également préconisée par Guski (1997).

Chapitre II

Les attributs de qualité

Selon Bech (1999), un “attribut sonore” est une caractéristique perceptive d'un stimulus sonore. Durant les dernières années, des études se sont focalisées sur la verbalisation de la perception sonore en construisant un lexique spécifique (Pedersen et Zacharov, 2008). Diverses méthodes d'élicitation ont été mises en oeuvre pour extraire les termes relatifs aux sons et plus récemment à la restitution sonore spatialisée. Les méthodes telles que ITU-R BS.1116 (1997) et ITU-R BS.1534 (2003) proposent une évaluation basée principalement sur la qualité audio de base. Il serait intéressant d'avoir plus d'informations sur les caractéristiques sonores dégradées par les systèmes de codage. Cependant, les attributs sonores sont nombreux et ne peuvent être tous inclus dans un test d'évaluation de qualité principalement pour des raisons protocolaires. L'étude décrite dans ce chapitre a pour but de réduire un ensemble d'attributs sonores à des catégories plus générales pouvant servir d'axes perceptifs d'évaluation de la qualité. Deux méthodes ont été mises en place : une méthode indirecte avec une analyse multidimensionnelle et une méthode directe avec une catégorisation libre suivie d'une analyse par cluster. Les catégorisations sont réalisées en français sans présenter de séquences sonores. Les résultats des deux méthodes sont comparés afin de définir des familles d'attributs sonores.

II.1 Les méthodes d'élicitation

Berg et Rumsey (1999) ont proposé la technique des grilles-répertoires (RGT : repertory grid technique) pour l'identification d'attributs spatiaux. Cette méthode se base sur les travaux de Kelly (1955) qui a développé la théorie des construits personnels. Un construit est une dimension, un concept choisi par le sujet pour différencier les événements qui lui sont proposés. L'utilisation de son propre langage permet au sujet d'être plus fiable lors de l'évaluation. Dans l'expérience de Berg (2006), la procédure de test se déroulait en trois étapes.

* La phase d'élicitation : trois stimuli étaient présentés à un auditeur. Il devait les comparer et indiquer les deux extraits qu'il percevait comme les plus similaires. Puis, il utilisait deux termes antagonistes pour décrire cette similarité perçue.
* La phase de notation : Les deux termes étaient ensuite considérés comme les extrémités d'une échelle bipolaire qui servait à évaluer les stimuli (figure II.1).

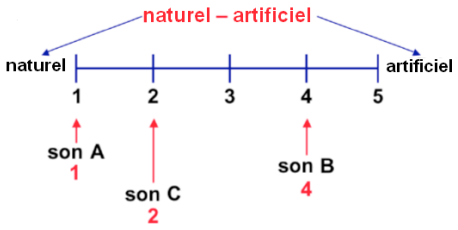


Fig. II.1 – Notation sur l'échelle de construits bipolaires (Berg, 2005).

Une grille-répertoire est construite avec les échelles et les notes pour chaque stimulus (figure II.2).

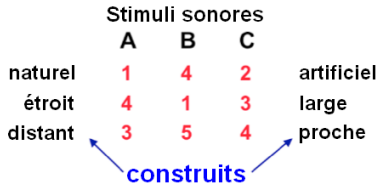


Fig. II.2 – Exemple de grille-répertoire pour un sujet (Berg, 2005).

* L'analyse des données : elle peut se faire avec une analyse par cluster ou une analyse en composantes principales. OPAQUE (Optimisation of Perceived Audio QUality Evaluation) est un logiciel qui réalise automatiquement une procédure RGT (repertory grid technique) et donc facilite l'élicitation et la notation (Berg, 2005). La RGT, grâce aux choix des termes par le sujets lui-même, permet de limiter l'influence du chercheur et de son propre référentiel (Cossette, 2004), notamment en ce qui concerne les échelles de notation imposées aux sujets. Un inconvénient de la RGT est que la différence entre deux extraits peut passer inaperçue lorsque ces deux sons sont toujours comparés avec un troisième son toujours plus différent que ces deux sons (Choisel et Wickelmaier, 2006).

Choisel et Wickelmaier (2006) ont développé l'analyse des structures perceptives (PSA : Perceptual Structure Analysis) pour extraire des attributs sonores d'extraits musicaux reproduits sur divers systèmes multicanaux (mono, stéréo, 5.0, ...). La méthode PSA se base sur des principes mathématiques : l'analyse de concepts formels (Ganter *et al.*, 1997) et la théorie des espaces de connaissance (Doignon et Falmagne, 1998). L'ensemble des extraits constitue un domaine noté X, et α est l'ensemble des sous groupes de X correspondant aux attributs sonores ; $\langle X, \alpha \rangle$ forme une structure perceptive (Choisel et Wickelmaier, 2006; Heller, 2000). Une représentation sous forme de treillis représente le domaine et ses sous-ensembles, les noeuds indiquent les caractéristiques partagées par les extraits sonores. Concernant le déroulement du test, il était demandé aux sujets d'écouter trois sons et de répondre par 'oui' ou 'non' à la question "Do sounds a and b share a feature which c does not have ?" (Est ce que les sons a et b partagent une caractéristique que c ne possède pas ?). Dans un premier temps, l'intérêt majeur est que l'identification des caractéristiques se fait sans demander aux sujets de les nommer. Ensuite, une fois les structures perceptives analysées, les auditeurs devaient nommer ces caractéristiques et en écrire une brève description.

Guastavino et Katz (2004) ont fait le choix d'une transcription verbale. Des paysages sonores ont été décrits via une tache de verbalisation libre. Une analyse linguistique

(lemmatisation, synonymie) a été réalisée sur les données collectées.

Une analyse multidimensionnelle (MDS : multidimensional scaling) a été utilisée par Susini *et al.* (1999) pour mettre en avant des dimensions perceptives sur les sons de voitures. Grey (1977) a également employé cette méthode pour révéler des dimensions du timbre musical. Lavandier *et al.* (2008a, 2005, 2008b) l'ont utilisé pour l'évaluation perceptive des enceintes acoustiques. Cette méthode sera détaillée dans la section II.5.

Toutes ces études ont permis d'établir une liste non exhaustive d'attributs pour qualifier le son, avec par exemple, la localisation, la distance, la couleur sonore, le bruit, la brillance, la sifflement, l'enveloppement, ... (tableau II.1).

Tab. II.1 – Synthèse d'études sur les attributs de qualité.

Nakayama *et al.* (1971)	Gabrielsson et Sjögren (1979)	Toole (1985)	Berg et Rumsey (2001)
Sensation of clearness	Clearness / Distinctness	Clarity definition	Localisation
	Brightness - Darkness	Brightness	Preference
	Sharpness/hardness - Softness	Softness	Envelopment
Sensation of fullness	Fullness - Thinness	Fullness	Width
	Feeling of space	Pleasantness	Presence
	Disturbing sounds	Hiss, noise, distortion	Naturalness
Depth of the image	Nearness	Impression of distance/depth	Source distance
	Loudness	Definition of sound image	Source width
		Continuity of the sound stage	Background noise level
		Fidelity	
		Abnormal effects	
		Reproduction of ambiance, spaciousness & reverberation	
		Perspective	
		Overall spatial rating	

Koivuniemi et Zacharov (2001)	Guastavino et Katz (2004)	Lorho (2005a)	Choisel et Wickelmaier (2006)
Tone Color	Coloration	Clarity	Clarity
Richness	Presence	Richness	Brightness
Hardness	Readability	Sense of distance	Spaciousness
Emphasis	Stability	Sense of direction	Envelopment
Naturalness	Naturalness/ Realism	Sense of movement	Naturalness
Sense of direction	Distance	Ratio of localisation	Elevation
Sense of depth	Localization	Quality of echo	Width
Sense of space	Spatial distribution of sound	Sense of space	Distance
Sense of movement	Spectral balance	Amount of echo	
Penetration		Balance of space	
Distance to events		Separability	
Broadness		Broadness	
		Distorsion	
		Disruption	
		Tone color	
		Balance of Sounds	

Les attributs hédoniques (plaisant, gênant...) font référence aux préférences d'une personne, à ce qu'elle aime ou n'aime pas (Bech, 1999). Les jugements hédoniques sont connus pour biaiser les tests d'écoute (Zielinski, 2006). Pour une meilleure évaluation d'extraits sonores, il faut être le plus objectif possible dans sa subjectivité. Si un auditeur a une préférence pour un extrait sonore, il ne doit pas pour autant l'évaluer systématiquement supérieur aux autres.

Les définitions des attributs portent à confusion et elles ne sont pas toujours identiques selon les études (Le Bagousse *et al.*, 2010). Par exemple, la définition de richesse par Koivuniemi et Zacharov (2001) est "the homogeneity of the timbre" et Lorho (2005a) décrit la richesse comme une "combination of harmonics and dynamics perceived in a sample". Cela peut donc engendrer des biais lors de tests d'évaluation. Cheminée et Dubois (2009) ont procédé à une analyse linguistique au sujet du vocabulaire utilisé sur la manière dont les pianistes parlent des sons de pianos. Il semble que les pianistes aient un vocabulaire commun. Cependant les adjectifs n'ont pas le même sens chez eux et dans un dictionnaire, parfois ils ne sont pas recensés, même s'ils paraissent compréhensibles comme par exemple le terme *percussif*. Ils ont souligné également que l'emploi du terme *clair* est différent selon le contexte. Il peut signifier : défini, sec/dur ou lumineux. Choisel et Wickelmaier (2007) ont défini la clarté ainsi : "plus le son est clair, plus il est possible de percevoir les détails de celui-ci". Pour Lorho (2005a), la clarté décrit "si le son parait clair ou sourd". Les définitions peuvent sembler similaires mais peuvent être interprétées différemment par les sujets. De plus, il est impossible d'inclure tous ces attributs (tableau II.1) dans un test d'évaluation de la qualité. Le test durerait trop longtemps et la complexité pour les testeurs serait trop importante. La définition de catégories plus générales mais représentant, chacune, un axe perceptif à part entière permettrait de répondre à ces problèmes.

II.2 Les catégorisations d'attributs existantes

L'unique critère évalué par les méthodes ITU-R BS.1116 et ITU-R BS.1534 est la qualité audio de base. Bien que ce soit un critère indispensable, il serait intéressant d'évaluer d'autres attributs dans des tests d'écoute pour connaître les axes de qualité dégradés par les codages audio. Pour les systèmes autre que monophoniques, ces deux méthodes proposent des attributs supplémentaires comme la qualité d'image stéréophonique ou encore la qualité frontale de l'image mais ils ne sont jamais utilisés en pratique.

Une autre méthode normalisée, l'EBU Tech 3286 Supp. 1 (2000) propose huit attributs généraux pour l'évaluation de qualité dédiée à la musique classique diffusée sur un système multicanal :

- La qualité de l'image frontale
- La qualité des arrières
- Impression spatiale
- Transparence
- Balance sonore
- Couleur sonore
- Absence de bruit et de distorsion
- Impression générale

D'autres catégorisations ont été mises en avant dans la littérature. Letowski (1989) a proposé le modèle MURAL (figure II.3). L'image auditive est divisée en deux attributs principaux : le timbre et l'espace ("spaciousness"). Successivement les catégories se subdivisent avec l'ajout d'attributs plus précis comme réverbération, clarté...

Fig. II.3 – Le modèle Mural de Letowski (1989).

Berg et Rumsey (2003) suggéraient un modèle générique pour la qualité audio qui inclurait la qualité timbrale, la qualité spatiale, la qualité technique et la qualité diverse qui contient ce qui ne figure pas dans les trois autres catégories (figure II.4).

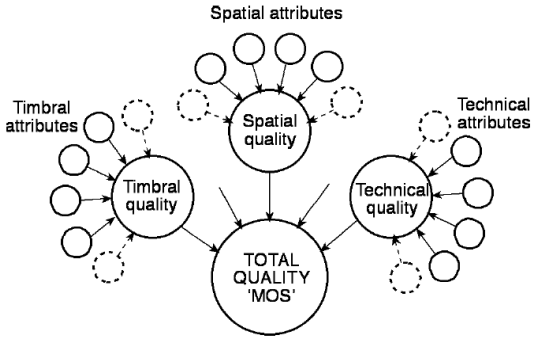


Fig. II.4 – La catégorisation proposée par Berg et Rumsey (2003).

Lorho (2005b) a trouvé trois catégories qui sont : l'aspect timbral, l'aspect spatial et l'accentuation des basses fréquences (timbral aspects, spatial aspects and low-frequency emphasis). Il a utilisé une méthode d'élicitation verbale où chaque sujet développait ses propres descripteurs. La méthode est nommée profil lexical individuel (IVP : individual vocabulary profiling). Une analyse par cluster a permis ensuite de former les catégories.

Berg et Rumsey (2000) ont utilisé une analyse par cluster pour catégoriser les termes relatifs à la dimension spatiale du son. Cette méthode d'analyse sera employée et expliquée dans le paragraphe 6.2.

Zielinski *et al.* (2005) ont défini trois attributs d'évaluation : la fidélité timbrale, la fidélité spatiale frontale et ambiophonique.

L'étude décrite dans ce chapitre a pour but de valider de manière objective certaines catégorisations énoncées précédemment. En utilisant deux méthodes, indirecte avec une MDS et direct avec une catégorisation libre suivie d'une analyse par cluster, des attributs sonores sont regroupés en catégories plus générales. Les catégorisations sont réalisées de manières sémantiques sans présenter de séquences sonores. Les tests sont réalisés en français. Les résultats des deux méthodes sont comparés afin de définir des familles d'attributs sonores.

II.3 Choix d'une liste d'attributs

En se basant sur une liste non exhaustive d'attributs perceptifs issus des études précitées (tableau II.1), 28 sont retenus dans cette étude et soumis à un groupe de sujets en vue de les catégoriser. Le tableau II.2 contient la liste des 28 attributs.

Tab. II.2 – Liste d'attributs proposée pour la catégorisation.

Attributs sonores		
Bourdonnement	Brillance	Bruit
Bruit de fond	Clarté	Coloration
Couleur du timbre	Coupure	Distance
Distorsion	Distribution spatiale du son	Dureté
Dynamique	Enveloppement	Equalisation
Fidélité	Homogénéité	Immersion
Largeur	Localisation	Précision
Profondeur	Réalisme	Réverbération
Richesse	Sifflement	Spatialisation
Stabilité		

Pour établir cette liste à partir du tableau II.1, les termes antonymes ont été exclus. Effectivement, ces termes peuvent générer des incohérences et être regroupés ensemble par certains sujets ou dans des catégories différentes par d'autres. Par exemple, les attributs "richesse/pauvreté" sont des termes opposés dans le sens mais ils caractérisent, tous deux, la même caractéristique sonore, "pauvreté" a donc été enlevé. Les attributs cités dans le tableau II.1 ont été soumis à 12 sujets "experts audio". Ceux peu énumérés par les différents auteurs et jugés non pertinents pour évaluer la qualité audio par au moins la moitié des sujets ont également été retirés (pénétration, lisibilité, authenticité...).

II.4 Les testeurs

18 testeurs sont recrutés pour réaliser les deux tests. Les auditeurs sont considérés comme "experts", conformément aux recommandations ITU-R BS.1116 (1997) et ITU-R BS.1534 (2003). Chaque sujet effectue les deux tests à environ une semaine d'intervalle. Cependant pour éviter l'influence d'un test sur l'autre, la moitié des personnes commence par le test de catégorisation libre tandis que l'autre moitié commence par la MDS.

II.5 Test A : L'analyse multidimensionnelle

La première méthode employée pour procéder à la catégorisation des attributs est une MDS (Borg et Groenen, 2005; Kruskal et Wish, 1978). Cette méthode indirecte permet de révéler des dimensions perceptives (Rumsey, 1998) par le biais d'une mesure de ressemblance entre des variables.

II.5.1 Description du protocole de test

Les testeurs doivent juger de la similarité/dissimilarité entre deux attributs issus de la liste décrite dans le tableau II.2. Les attributs sont présentés par paire comme le montre la figure II.5 qui représente l'interface conçue pour le test. L'échelle de notation est continue de 0 à 1, avec respectivement pour extrémités les termes "très différents" et "très similaires" (Susini *et al.*, 1999). Chaque testeur juge de la similarité perçue entre deux termes en positionnant le curseur à la position désirée sur l'échelle puis clique sur "suivant" pour accéder à la paire suivante (annexe A.1). Il ne peut pas revenir en arrière. C'est un test lexical, aucun son n'est présenté. Les testeurs se basent sur leur propre interprétation des attributs (ils ne sont pas autorisés à consulter leurs définitions dans un dictionnaire).

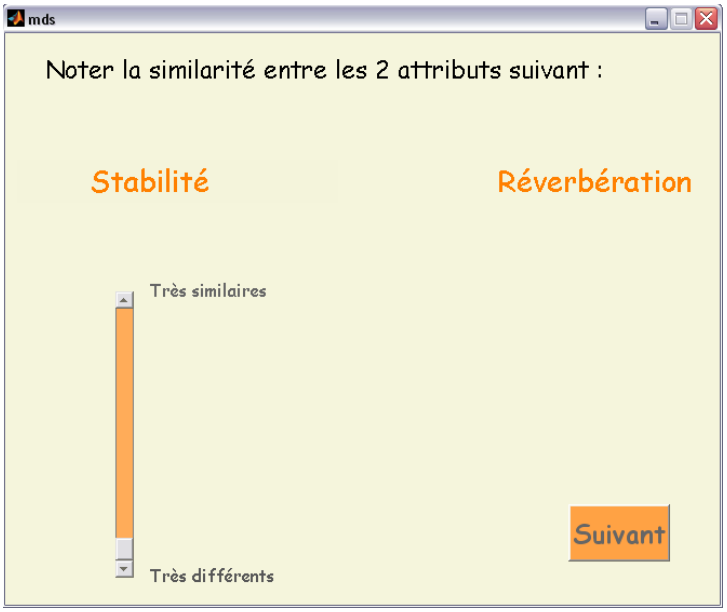


Fig. II.5 – Interface de test pour la MDS.

Dans un premier temps, les testeurs sont soumis pendant environ 5 minutes à une phase de familiarisation de l'interface et des attributs à comparer. Le test principal contient 28 attributs. Toutes les paires possibles sont comparées. Donc au total $\frac{n*(n-1)}{2}$, soit 378 paires sont évaluées. Le test dure approximativement une heure. Les testeurs peuvent faire des pauses quand ils le désirent, au moins une pause est recommandée.

Les notes comprises entre 0 et 1, attribuées par chaque testeur pour une paire d'attributs perceptifs, sont enregistrées dans des matrices carrées symétriques (28*28).

II.5.2 Résultats de l'analyse multidimensionnelle

Les valeurs obtenues entre les différents attributs correspondent à des mesures de distance, dans le but de construire une représentation géométrique multidimensionnelle, dit espace perceptif (Yannou et Deshayes, 2006).

Le modèle INDSCAL (Individual differences scaling) proposé par Carroll et Chang (1970) est une MDS pondérée. Ce modèle est choisi car il prend en compte les différences inter-individuelles c'est-à-dire le poids que les sujets attribuent aux dimensions (Etame, 2008). La MDS non métrique (Kruskal, 1964), contrairement à la MDS métrique, privilégie l'ordre des proximités plutôt que leurs valeurs exactes pour construire l'espace perceptif (Borg et Groenen, 2005). À partir des distances observées d, des disparités $f(d)$ (écarts entre valeurs d'une variable) sont calculées, vérifiant une règle de monotonie si $d1 < d2$ alors $f(d1) <= f(d2)$. L'analyse des données est réalisée avec le logiciel SPSS par la procédure INDSCAL non métrique (Takane *et al.*, 1977).

Deux paramètres principaux résultant de l'analyse réalisée permettent de déterminer le nombre de dimensions optimal pour la représentation de l'espace perceptif. Ces deux indices sont le Stress et le RSQ. Le Stress illustre par un processus itératif l'ajustement entre les distances dans la représentation graphique et les disparités des distances observées en minimisant leur différence. Le RSQ est la proportion de variance, le coefficient de corrélation au carré (Naes et Risvik, 1996).

Il a été énoncé par Tournois et Dickes (1993) que le nombre de dimensions approprié est obtenu lorsque l'ajout d'une dimension supplémentaire ne réduit que faiblement la valeur du stress. De ce fait et au vu des valeurs indiquées dans le tableau II.3, l'espace contenant 5 dimensions semble être le plus adéquat (Stress = 0.20, RSQ=0.43).

Tab. II.3 – Valeurs du stress et du RSQ.

Nbr de dim.	2	3	4	5	6
Stress	0.3851	0.29884	0.23757	**0.20091**	0.18118
RSQ	0.38228	0.40062	0.41849	**0.43363**	0.43599

Effectivement, sur la courbe des valeurs du RSQ (figure II.6), il y a bien une cassure à partir de la dimension 5. Cette même observation peut être faite sur la courbe du stress mais de manière moins évidente.

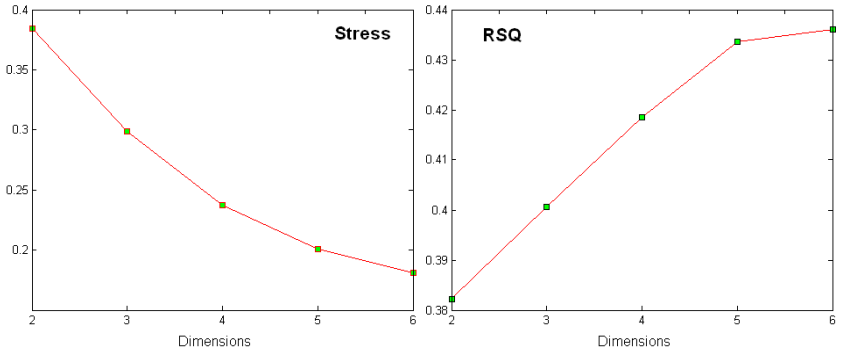


Fig. II.6 – Courbe du stress et du RSQ.

Le tableau II.4 répertorie le poids de chacune des dimensions de l'espace. Il est à noter que la dimension 1 et la dimension 2 sont largement plus importantes que les autres.

Tab. II.4 – Poids des dimensions, considérant l'espace perceptif à 5 dimensions.

Dimension	1	2	3	4	5
Poids	0.1550	0.1040	0.0699	0.0530	0.0517

Une dimension est un axe perceptif et tous les attributs sont répartis sur cet axe. Ici, le but est de trouver une catégorisation des attributs (un attribut appartient à une seule catégorie) plutôt que l'interprétation des dimensions. La figure II.7 représente la distribution des attributs en fonction des dimensions 1 et 2 en considérant les coordonnées dans l'espace à 5 dimensions.

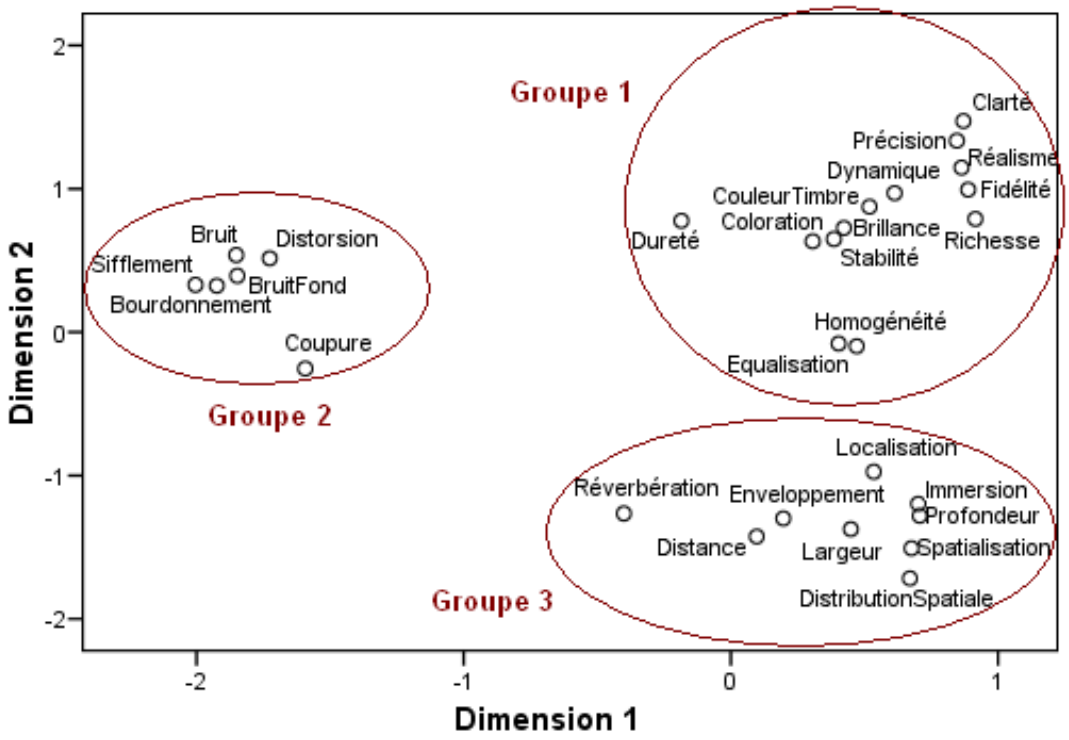


Fig. II.7 – Espace perceptif, dimension 1 et 2.

Trois groupes d'attributs se distinguent clairement. Ils sont détaillés dans le tableau II.5. Les attributs "homogénéité" et "équalisation" semblent être associés au groupe 1 bien que la catégorisation soit moins évidente pour ces deux termes.

La figure II.8 montre la projection des attributs en fonction cette fois des dimensions 2 et 3.

Tab. II.5 – Classification des attributs selon les dimensions 1 et 2.

Groupe 1	**Groupe 2**	**Groupe 3**
Fidélité	Bruit de fond	Réverbération
Dureté	Bruit	Spatialisation
Richesse	Distorsion	Distribution spatiale du son
Homogénéité	Coupure	Localisation
Précision	Sifflement	Largeur
Couleur du timbre	Bourdonnement	Distance
Coloration		Enveloppement
Brillance		Profondeur
Clarté		Immersion
Dynamique		
Réalisme		
Stabilité		
Equalisation		

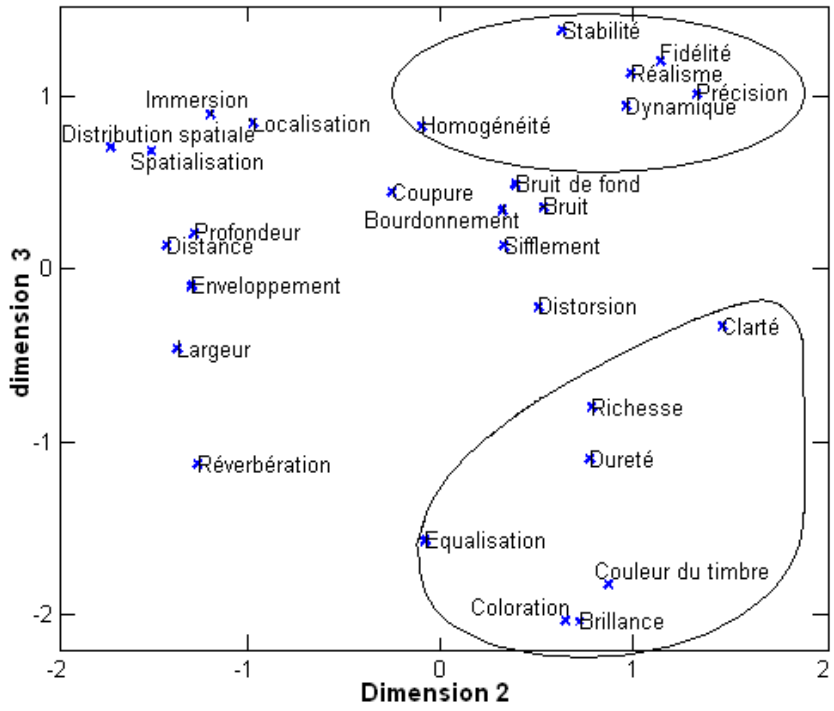


Fig. II.8 – Espace perceptif, dimension 2 et 3.

Il est à remarquer que le groupe 1 se scinde en deux sous-groupes comprenant les attributs suivants :

1. Richesse, brillance, dureté, couleur du timbre, coloration, clarté et équalisation
2. Réalisme, fidélité, stabilité, homogénéité, dynamique et précision

Les deux termes “homogénéité” et “équalisation” trouvent bien leur place dans ces sous groupes, ce qui confirme leur appartenance au groupe 1 du tableau II.5.

L'analyse des dimensions 4 et 5 ne permet pas d'affiner cette catégorisation. En résumé par la méthode MDS, trois catégories d'attributs sont révélées dont une séparable en deux sous catégories.

II.6 Test B : La catégorisation libre

II.6.1 Procédure de test

Pour ce second test, une méthode directe, une catégorisation libre des attributs est employée. Les sujets utilisent une interface dotée de la méthode du glisser-déposer pour composer librement leurs catégories, en déplaçant les attributs d'une colonne à une autre (figure II.9).

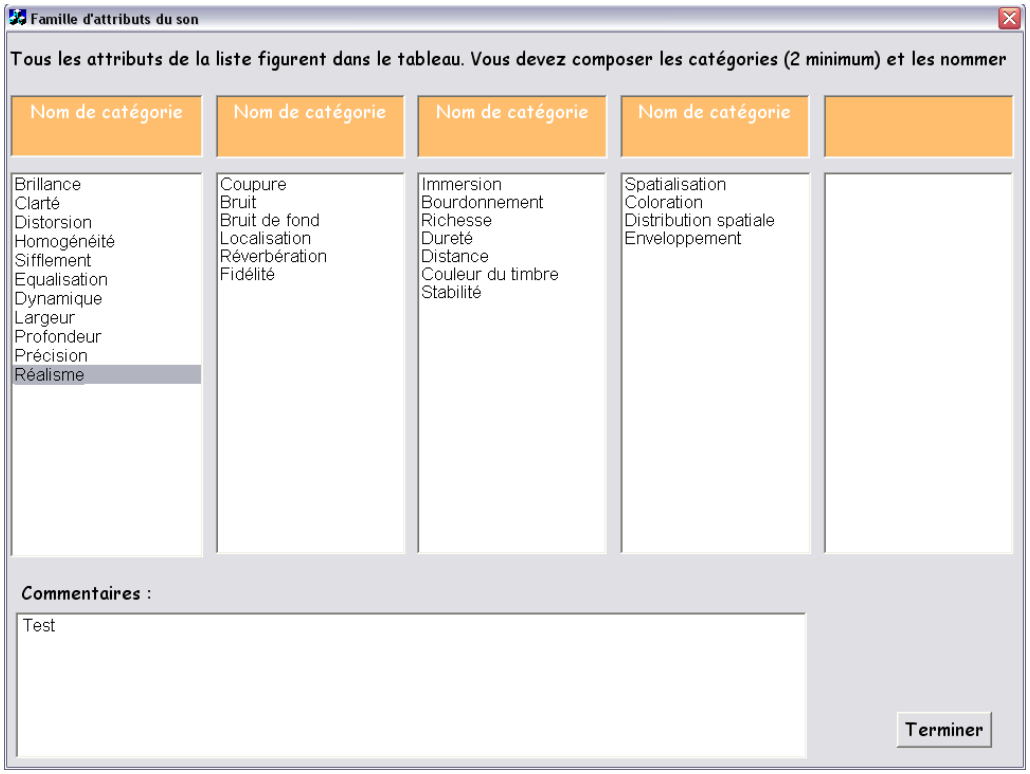


Fig. II.9 – Interface de test pour la catégorisation libre.

La liste est constituée de 28 attributs identiques à l'expérience A (tableau II.2). La seule consigne donnée (annexe A.2) est de constituer au minimum deux et au maximum cinq catégories avec tous les attributs (aucun attribut ne peut être laissé de coté). Le choix d'imposer un maximum est de limiter le nombre de catégories en vu de les inclure dans un test d'écoute comme axes d'évaluation. Pour terminer, les sujets doivent nommer les catégories qu'ils ont eux-mêmes formées. Comme pour l'expérience A, les définitions des attributs ne sont pas accessibles et les testeurs doivent se fier à leur propre interprétation.

II.6.2 Analyse par cluster

L'analyse par cluster est une méthode de regroupement d'objets ou groupes d'objets par des calculs de distances. Dans cette étude la méthode utilisée est une classification

ascendante hiérarchique (CAH). Les éléments vont au fur et à mesure fusionner avec les points les plus proches, au sens de la distance choisie. Le dendrogramme est le diagramme généré suite à une CAH. Il permet de visualiser les groupes formés.

La méthode appliquée est celle de Ward (1963). Son principe se base sur le calcul de l'inertie c'est-à-dire la distance au carré entre les centres de gravité des classes pondérées par les effectifs de ces classes. Une classe est un attribut sonore ou un groupe d'attributs. Le but est de minimiser l'inertie totale, de façon à ce que l'augmentation de l'inertie intraclasse soit minimum lors de la fusion de deux clusters (Théorème de Huygens, illustré par la figure II.10).

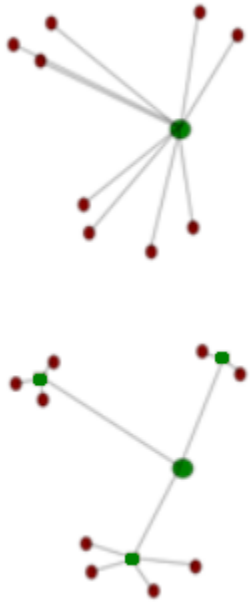


Fig. II.10 – Illustration du théorème de Huygens.

L'analyse est réalisée avec la fonction Clustering incluse dans Matlab et ci-dessous la formule de distance utilisée :

$$d^2(r,s) = n_r n_s \frac{||\bar{x}_r - \bar{x}_s||_2{}^2}{n_r + n_s} \; avec \; \bar{x}_r = \frac{1}{n_r} \sum_{i=1}^{n} x_{ri}$$

où $||\ ||_2$ est la distance euclidienne, $\bar{x}_r$ et $\bar{x}_s$ sont les centroïdes des clusters r et s, et n_r et n_s sont les nombres d'éléments dans les clusters r et s.

Pour ce test, 18 tableaux sont construits, chacun contenant les catégories composées par chacun des 18 testeurs. 9 sujets sur les 18 forment 4 catégories, 5 d'entres eux constituent 5 catégories et les 4 restants créent 3 catégories, soit au total 73 catégories d'attributs. Une matrice (28*73) est construite contenant d'un coté la liste d'attributs et de l'autre toutes les catégories formées. Pour chiffrer cette matrice, la note d'une valeur de 1 est attribuée à un attribut lorsqu'il figure dans une catégorie et 0 sinon (figure II.11). La méthode de Ward est appliquée sur cette matrice.

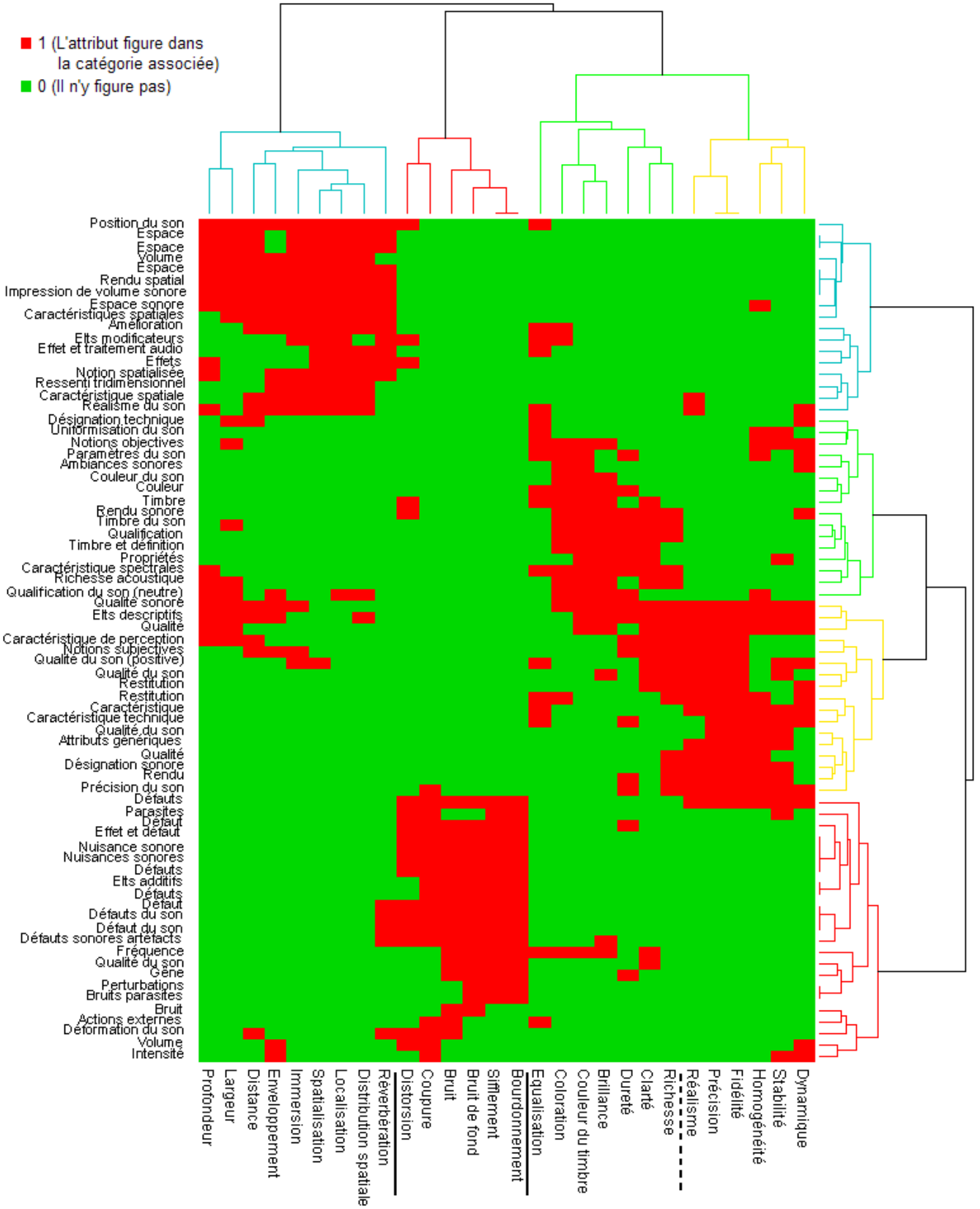


Fig. II.11 – Dendrogramme issu de la catégorisation libre.

II.6.3 Résultats du dendrogramme

La figure II.12 résume les résultats de l'analyse par cluster pour le regroupement des attributs. Trois groupes principaux sont formés. Le groupe 1 se divise en deux sous-groupes a et b.

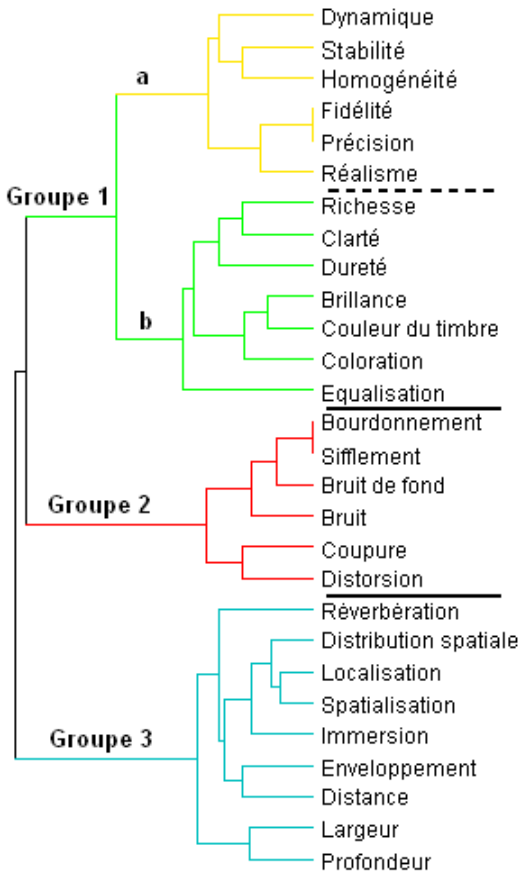


Fig. II.12 – Classification des attributs par l'analyse en cluster.

Concernant l'attribution du nom de chaque catégorie, le dendrogramme permet également d'associer les noms donnés par les testeurs aux catégories d'attributs formées (tableau II.6).

Pour chaque groupe, le nom choisi correspond à l'appellation ayant obtenu le plus grand nombre d'occurrences, c'est-à-dire lorsqu'un sujet a employé ce terme. Ainsi avec 9 occurrences, le groupe 2 est identifié comme "Défauts". Le groupe 3 qui fait référence à l'aspect spatial du son est nommé "Espace". Le groupe 1 se divise en deux sous-groupes l'un se rapportant au "Timbre" et l'autre à des caractéristiques plus générales.

Tab. II.6 – Noms des familles d'attributs donnés par les testeurs, le chiffre entre () indique le nombre d'occurrences.

Groupe 1a	**Groupe 1b**	**Groupe 2**	**Groupe 3**
Qualités (6)	Désignation technique	Défauts (9)	Position du son
Eléments descriptifs	Uniformisation du son	Parasites	Espace (4)
Caractéristiques de perception	Notions objectives	Effets	Volume
Paramètres du son	Notions subjectives	Nuisances sonores	Rendu spatial
Ambiances sonores	Restitution (2)	Eléments additifs	Impression de volume
Caractéristique (2)	Couleur (2)	Artefacts	Caractéristiques spatiales (2)
Attributs génériques	Timbre (3)	Fréquences	Améliorations
Désignation sonore	Rendu sonore	Qualité du son	Eléments modificateurs
Rendu	Qualification (2)	Gêne	Effets et traitement audio
Précision du son	Propriétés	Perturbations	Effets
	Caractéristiques spectrales	Bruit parasites	Notions spatialisées
	Richesse acoustique	Bruit	Ressenti tridimensionnel
		Actions externes	Réalisme
		Déformation du son	Immersion
		Volume	
		Intensité	

II.7 Discussion

Les deux méthodes mises en place, la MDS (méthode indirecte) et l'analyse en clusters de la catégorisation libre (méthode directe), donnent le même résultat. Les catégorisations faites dans cette étude révèlent trois principales familles d'attributs dont l'une d'entre elles se scinde en deux sous-groupes. De ce fait, trois ou quatre catégories peuvent être prises en considération. La catégorisation libre faite par les testeurs a permis d'associer un nom aux catégories : une est associée à l'*Espace*, une seconde fait référence aux *Défauts* perçus et la troisième porte sur la *Qualité*. Cette dernière se divise en deux groupes. Ainsi, les quatre catégories révélées dans cette étude sont :

- Défauts : ce sont les parasites ou nuisances présents dans le son, e.g. bruit, distorsion, bruit de fond, sifflement, bourdonnement, coupure...
- Espace : réfère à l'impression spatiale relative aux caractéristiques spatiales, e.g. profondeur, largeur, localisation, distribution spatiale, réverbération, spatialisation, distance, enveloppement, immersion...
- Timbre : se rapporte à la couleur sonore, e.g. brillance, couleur du timbre, coloration, clarté, dureté, équalisation, richesse...
- Qualité : composé de homogénéité, stabilité, précision, réalisme, fidélité, dynamique...

La catégorie *Qualité* regroupe des attributs qui ne se semblent pas homogènes. Elle peut être comparée à la *qualité diverse* proposé par Berg et Rumsey (2003) qui regroupe les caractéristiques restantes. Ci-dessous les quatre catégories qu'ils suggéraient :

- qualité timbrale : réfère à la couleur sonore, le timbre.
- qualité spatiale : réfère la nature tri-dimensionnelle des sources sonores et des environnements.
- qualité technique : réfère aux distorsions, sifflements, bourdonnements etc.
- qualité diverse : réfère aux propriétés restantes.

Les trois autres catégories, qualité timbrale, spatiale et technique sont similaires aux catégories nommées *Timbre*, *Espace*, *Défauts* découvertes dans ce chapitre et confirme donc cette classification.

Les catégories mises en avant pourront être incluses dans un test d'écoute et leurs pertinences étudiées en tant qu'axes d'évaluation pour la qualité audio. Il sera alors intéressant de vérifier l'intérêt d'un test multi-critères et de mesurer le poids relatif de chaque attribut par rapport à l'évaluation de la qualité globale. En ce sens, la catégorie *Qualité* peut être un biais important. Son intitulé est ambigu et peut être confondu avec la qualité globale par les auditeurs. De plus, les attributs constituant cette catégorie réfèrent à des caractéristiques générales du son.

La plupart des études qui utilisent des attributs proposent généralement deux catégories : le timbre et l'espace (par exemple Letowski (1989)). Ces deux attributs sont présents dans toutes les catégorisations. Pour l'évaluation sur des systèmes audio 5.1, Zielinski *et al.* (2005) ont séparé l'espace en deux attributs : la fidélité spatiale frontale et la fidélité spatiale ambiophonique (basés sur les attributs proposés par l'ITU). En revanche, cette catégorisation n'est valable que pour ce système de restitution. Le terme "fidélité" a été employé car les tests audio incluaient une référence explicite. Ces attributs ont, ensuite, été inclus dans des tests d'écoute. Ainsi, Rumsey *et al.* (2005) et Marins *et al.* (2008) ont montré que la fidélité timbrale avait plus d'influence sur la qualité audio de base que la fidélité spatiale.

Les autres catégorisations proposaient, pour la plupart, un nombre de catégories élevé qui rend un test difficile à mener, comportant une phase de familiarisation importante et donc une durée de test très longue.

La catégorie portant sur les défauts a peu été mise en avant dans les études passées. Pourtant, dans le contexte du codage audio, notamment pour les dégradations engendrées, cet attribut est particulièrement pertinent. Marins *et al.* (2006, 2007) ont procédé à des tests d'écoute en proposant quatre attributs liés à des artefacts de codages : limitation de bande, bruits d'oiseaux ("birdies"), étalement temporel, distorsions spatiales. Ces quatre dégradations ont été évaluées et comparées à la qualité audio de base (BAQ). La limitation de bande et l'étalement temporel sont les artefacts qui affectaient le plus la BAQ. Pour l'évaluation de qualité, il sera intéressant de vérifier l'influence de la catégorie *Défauts* sur la qualité globale en comparaison du *Timbre* et de l'*Espace*.

II.8 Conclusion

La définition d'un attribut prête parfois à confusion et l'utilisation d'une catégorie plus globale peut permettre de contourner ce biais. Cette étude a permis de catégoriser une liste d'attributs en familles plus générales et moins nombreuses dans le but de les inclure comme axes de qualité perçue dans un test d'écoute. Deux méthodes ont été utilisées : une MDS et une catégorisation libre étudiée avec une analyse par cluster. Les deux méthodes obtiennent les mêmes résultats avec la composition de 3 catégories, une portant sur les *Défauts*, l'autre sur l'*Espace* et une troisième qui se divise en deux dont l'une est le *Timbre*.

Ces trois attributs seront inclus dans des tests d'écoute dédiés à l'audio spatialisé et

donc évalués par des auditeurs. Il sera alors intéressant de vérifier l'intérêt d'une évaluation multicritère mais également de mesurer le poids de chacune de ces catégories sur l'évaluation de la qualité globale.

Chapitre III

La présentation des attributs de qualité dans un test d'écoute

Quatre catégories d'attributs ont été définies dans le chapitre II. La catégorie *Qualité* est constituée d'attributs qui semblent non homogènes. De plus, elle peut être assimilée à la qualité globale ou qualité audio de base (BAQ) du fait de la proximité lexicale de leurs noms. Pour éviter toute confusion, cette catégorie n'est donc pas prise en compte dans la suite des travaux réalisés. Pour simplifier la lecture, le terme "attribut" remplace le terme "catégorie d'attributs" dans la suite du texte. Les trois attributs inclus dans le test d'écoute décrit dans ce chapitre sont donc le *Timbre*, l'*Espace* et les *Défauts*. Une question se pose sur la manière de proposer ces attributs durant un test d'écoute. Les études ont tendance à affirmer que chaque attribut doit être évalué indépendamment (Rumsey *et al.*, 2005; Lorho, 2010). La recommandation ITU-R BS.1534, explique que les testeurs "peuvent se sentir perdus et/ou déconcertés, ayant à répondre à de multiples questions relatives à un stimulus donné". Pour vérifier cela, l'étude réalisée ici vise à tester deux types de présentation : l'évaluation de chaque attribut indépendamment et l'évaluation des trois attributs simultanément. De plus, le but d'une évaluation multicritère est d'identifier les caractéristiques altérées par les dégradations proposées. Ainsi, la corrélation entre les attributs et la qualité globale est calculée et, par le biais d'une régression linéaire, un poids est attribué à chacun des trois attributs pour quantifier son influence sur la qualité globale.

III.1 Protocole expérimental

III.1.1 Conditions d'écoute

Le test est réalisé dans une salle d'écoute qui respecte les conditions de la recommandation ITU-R BS.1116. La pièce est de forme cubique et la surface est de vingt mètres carrés. La hauteur est de deux mètres cinquante. Le système de restitution sonore est un système 5.1 placé selon la recommandation ITU-R BS.775-2. Il se compose de cinq enceintes *Genelec 8040A* disposées sur un cercle de quatre mètres de diamètre et d'un caisson de basse *Genelec 7070A*. La carte son est une carte audio Haute Résolution *Digigram VX 882 HR*. La chaîne audio est constituée d'un convertisseur numérique/analogique *Apogee Rosetta 800* et d'un moniteur *SPL 2380S*. Ce moniteur permet la gestion des sources sonores 5.1 et le contrôle du volume sonore. Les auditeurs sont libres de régler le niveau sonore à leur convenance (Koehl et Paquier, 2013).

III.1.2 Sujets

Le panel de testeurs est composé de vingt-quatre personnes qualifiées d'"expert" au sens des normes ITU. Ces personnes ont l'habitude de participer à des tests d'écoute subjectifs et travaillent pour la plupart en lien avec le domaine audio ou sont musiciens.

III.1.3 Séquences sonores

Le choix des extraits présentés aux auditeurs est une des premières étapes à réaliser dans la conception d'un test d'écoute.

Six extraits audio sont soumis à évaluation. Ces extraits couvrent plusieurs types de contenus : musique, sons environnementaux, scènes de films. Les séquences durent en moyenne vingt secondes. Le tableau III.1 décrit les extraits sonores.

Tab. III.1 – Description des extraits sonores.

Nom	*Description*	*Durée (s)*
Fight	Une bagarre entre deux personnes avec explosions	22.5
Foot	Match de foot, commentaires au centre et ambiance de foule sur les arrières	17
Mer	Son de vague en bord de mer	20.7
Orchestre	Orchestre philharmonique	20.3
Jazz	Musique jazz, guitares, saxophone	19.4
Milanof	Sons divers, enfants, tonnerre et une voix tournante	20.2

Pour chacun de ces six extraits, six versions, plus ou moins dégradées, sont évaluées comprenant la version originale, trois ancres, une spécifique à chaque attribut et deux codages audio (tableau III.2), ce qui fait un total de trente-six séquences à évaluer.

Tab. III.2 – Description des versions évaluées.

Nom	*Description*
Original	Le fichier original (non dégradé)
ancreT	Filtrage à 3.5 kHz
ancreD	Ajout de bruit rose
ancreS	Inversion canal avant droit R et arrière gauche Ls
HEAAC-160	HE-AAC à 160 kbits/s
MP3-128	MP3 à 128 kbits/s

Les codages MP3 et HEAAC figurent parmi les plus utilisés et sont donc choisis pour ce test.

Le MP3 pour MPEG-1/2 Audio Layer 3 (ISO/IEC 11172-3, 1993) est l'algorithme de compression audio le plus répandu. Il est utilisé par une majorité des plates-formes de téléchargement de musique, par les constructeurs d'appareils électroniques (lecteurs DVD-CD-DivX, baladeurs...) et également pour les jeux PC. Sa popularité auprès des consommateurs en a fait un format de prédilection. Son principe se base sur un modèle

psychoacoustique exploitant l'effet de masque. C'est un système "destructif" qui supprime des informations dans le spectre en commençant par les fréquences peu perceptibles (haut débit) jusqu'à des altérations très marquées (bas débit) engendrant des bruits parasites, des pertes en haute fréquence (Brandenburg, 1999).

L'AAC (MPEG-2 Advanced Audio Coding) (ISO/IEC 13818-7, 2006) a été développé dans le but de succéder au format MP3. Il a été adopté par Apple et par Nintendo entre autre. Ce codage offre un meilleur ratio qualité/débit que le MP3. Selon le profil AAC et l'encodeur MP3, un fichier AAC encodé à 96 kbits/s est équivalent ou meilleur en terme de qualité qu'un fichier MP3 à 128 kbits/s (Autti et Biström, 2004). Le profil le plus déployé est l'AAC-LC (Low Complexity), notamment pour les applications bas-débit comme le streaming audio. L'HE-AAC (MPEG4 High Efficiency AAC) est une extension de l'AAC-LC (Low Complexity) complété de l'outil de reconstruction de bande spectrale (SBR : Spectral Band Replication) (Wolters *et al.*, 2003; ISO/IEC 14496-3, 2005). Ce format est utilisé sur les décodeurs TNT et pour la radio numérique.

Trois attributs de qualité sont évalués, par conséquent, trois ancrages sont inclus dans le test d'écoute. Chacun est spécifique à un des trois attributs (*Timbre, Espace* ou *Défauts*), c'est-à-dire qu'il dégrade uniquement l'attribut auquel il est associé. L'ancrage timbral "ancreT" correspond à un filtrage passe-bas butterworth d'ordre 8 à 3.5 kHz de l'original. L'ancrage spécifique à l'attribut *Défauts* "ancreD" consiste à ajouter du bruit rose sur chacun des canaux (RSB $\approx$ 30dB). Pour concevoir l'ancrage spatial "ancreS", une inversion est réalisée entre le canal avant droit R et le canal arrière gauche Ls afin de créer une incohérence spatiale.

III.1.4 Déroulement du test

Le test se déroule en deux sessions, la première consiste à évaluer la qualité globale et la seconde consiste à évaluer les trois attributs : *Timbre, Espace* et *Défauts*. La méthode de test s'inspire de la méthode MUSHRA (ITU-R BS.1534, 2003). Toutes les versions d'un même extrait sont présentées simultanément.

L'expression "qualité globale" est employée au lieu de "qualité audio de base" (BAQ). Pour rappel, la définition de la BAQ implique une comparaison directe entre une référence explicite et les différentes versions évaluées. Or, une référence explicite impose la qualité maximale puisqu'elle doit être évaluée au maximum de l'échelle de notation (ITU-R BS.1534, 2003). Le but, ici, est de laisser libre la notation en haute qualité et également de s'exempter d'une tache de reconnaissance systématique de cette référence et de comparaison des objets du test avec celle-ci. Ceci permet, ainsi, de privilégier une évaluation de qualité plutôt que de fidélité. Le test d'écoute réalisé dans ce chapitre ne présente donc pas de référence explicite. Cependant, la version originale (non dégradée) est incluse dans le test comme une référence implicite haute qualité mais pas forcément maximale.

L'échelle de notation est une échelle continue de qualité. Elle ne comporte pas de labels intermédiaires, sources de biais pour les auditeurs (chapitre I section 4.4). Pour éviter ces biais, Marins *et al.* (2008) utilisent les termes "basse fidélité" et "haute fidélité" pour nommer les extrémités de leur échelle. Puisque le test réalisé ici ne présente pas de référence explicite, le terme "fidélité" ne peut pas être employé. Les extrémités de l'échelle sont donc nommées "Basse qualité" et "Haute qualité". Les valeurs des notes ne sont pas visibles par

l'utilisateur. Une consigne est donnée aux auditeurs : ils doivent noter au maximum de l'échelle, le stimulus qu'ils perçoivent comme étant de plus haute qualité et ce, pour chaque attribut évalué. Cette consigne est obligatoirement respecté. En effet, si non, un message de rappel s'affiche et l'auditeur pourra passer à l'extrait suivant que lorsqu'il aura attribué la note maximale à une des versions.

Dans un premier temps, tous les auditeurs évaluent la qualité globale du son. La figure III.1 représente l'interface d'évaluation développée avec le logiciel Matlab©.

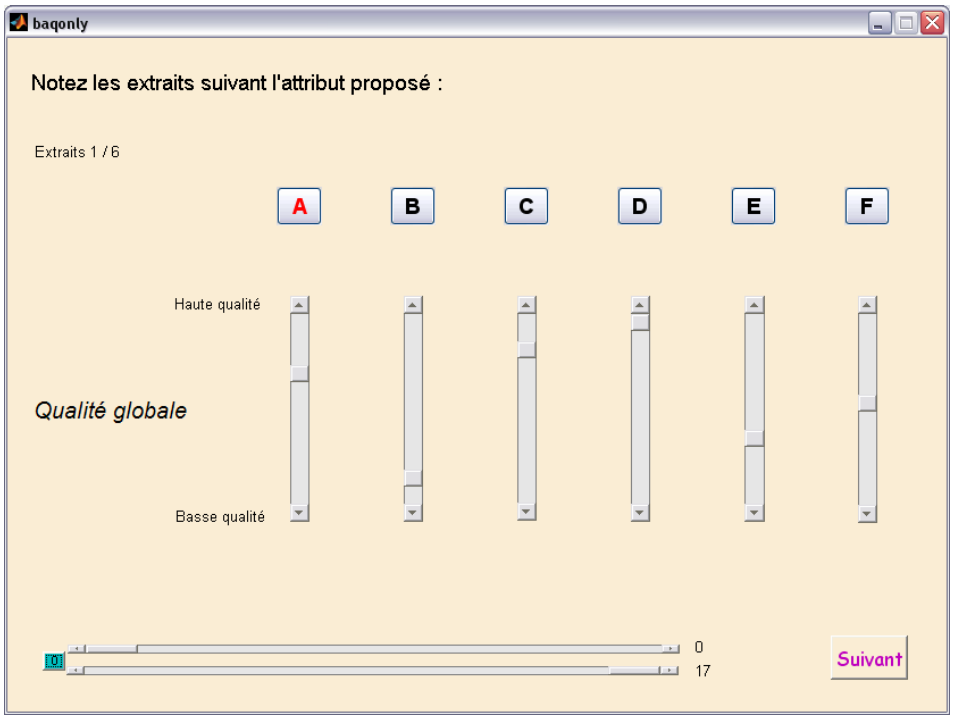
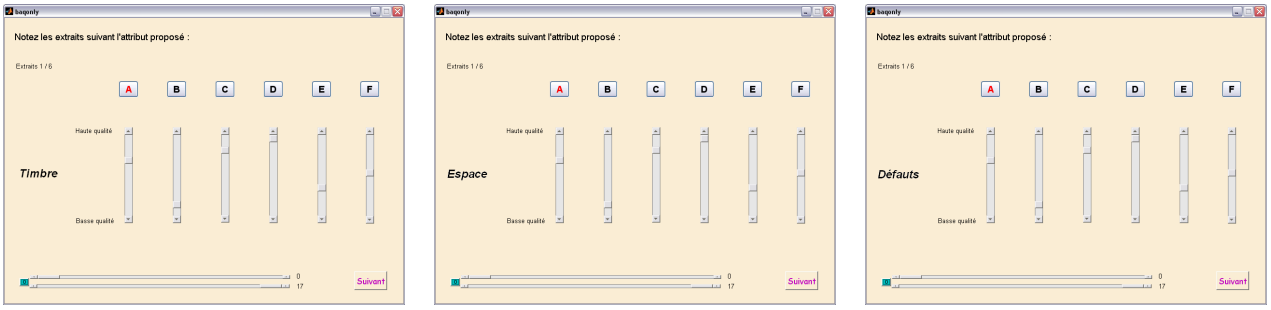


Fig. III.1 – Interface de test utilisée pour l'évaluation de la qualité globale.

Les séquences (extraits et versions) sont distribuées dans un ordre aléatoire pour chaque auditeur afin d'éviter une identification des dégradations et de minimiser l'effet de fatigue qui peut apparaître en fin de test. Pour chaque extrait proposé, les auditeurs écoutent les six versions en sélectionnant les lettres A, B, C, D, E ou F. Les séquences peuvent être écoutées autant de fois que nécessaire. Les deux curseurs horizontaux, de par leurs positions, permettent aux auditeurs de concentrer leur écoute sur un passage réduit de l'extrait. Le curseur supérieur permet de choisir l'instant où doit démarrer la séquence et le curseur inférieur désigne le temps auquel la lecture doit s'arrêter. Les auditeurs peuvent passer d'une version à l'autre en cours de lecture en sélectionnant les lettres A, B, C... Ensuite, pour évaluer la qualité perçue, ils positionnent le curseur de l'échelle de qualité au niveau voulu, compris entre basse et haute qualité, pour chaque version. Les auditeurs ont l'obligation de noter une des versions au maximum de l'échelle. Une fois satisfaits de leur notation, ils cliquent sur le bouton "Suivant" pour accéder à l'extrait suivant ; après quoi il n'est plus possible de revenir en arrière. Les consignes de test données aux auditeurs sont détaillées dans l'annexe A.3 et A.4.

Dans la seconde partie du test, qui concerne l'évaluation des attributs de qualité, le panel est divisé en deux groupes. Onze auditeurs évaluent les trois attributs séparément dans trois sessions successives (figure III.2). Les attributs sont proposés dans un ordre aléatoire pour chaque auditeur.



(a) Evaluation du timbre. (b) Evaluation de l'espace. (c) Evaluation des défauts.

Fig. III.2 – Interfaces de test utilisées pour l'évaluation de chaque attribut.

Le second groupe, composé de treize auditeurs, évaluent les trois attributs simultanément sur une même interface (figure III.3).

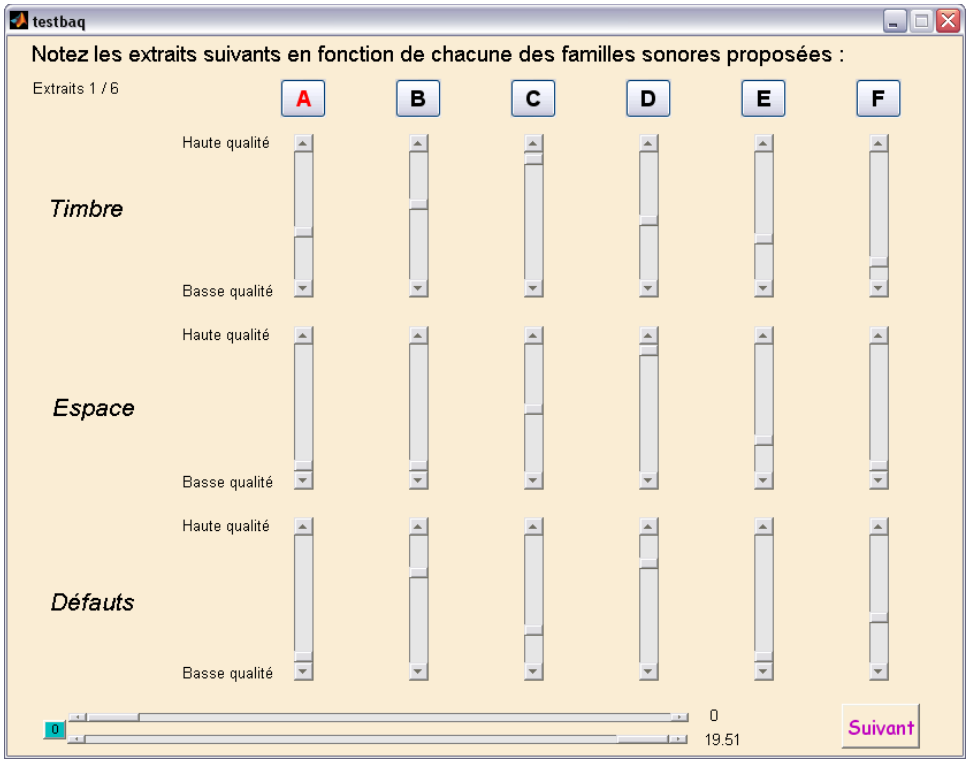


Fig. III.3 – Interface de test utilisée pour la présentation simultanée des attributs Timbre, Espace, Défauts.

III.2 Résultats

Il est important de rappeler que notre intérêt se porte sur la méthode d'évaluation et non sur le niveau de qualité des codeurs. L'échelle de notation "basse qualité" et "haute qualité" est transposée sur une échelle numérique comprise entre 0 et 1 avec précision de 10^{-2}.

III.2.1 Comparaison des deux modes de présentation des attributs

La première information relevée est le temps nécessaire aux auditeurs pour réaliser le test : la présentation successive des attributs a pris en moyenne 73 minutes alors que la présentation simultanée a pris environ 53 minutes. Ce gain de temps permet de limiter l'influence de la fatigue des auditeurs. Schatz *et al.* (2012) recommandent de ne pas excéder une durée de test de 90 minutes bien que les résultats semblent rester fiables malgré l'apparition de signes de fatigue. Notons cependant que les tests en psychoacoustique ne dépassent que rarement 1h.

Les tests d'hypothèse sont réalisés avec le logiciel *Statistica*. Les résultats obtenus pour chaque groupe d'auditeurs ont été comparés. Un test de Student a permis de montrer que la méthode de présentation des attributs n'a pas d'influence significative sur la notation des auditeurs. La figure III.4 montre les moyennes ainsi que les intervalles de confiance à 95% obtenus pour chaque version évaluée (codages et ancrages), moyennée sur toutes les séquences sur les deux groupes d'auditeurs (Groupe 1 : présentation successive des attributs ; Groupe 2 : présentation simultanée des attributs).

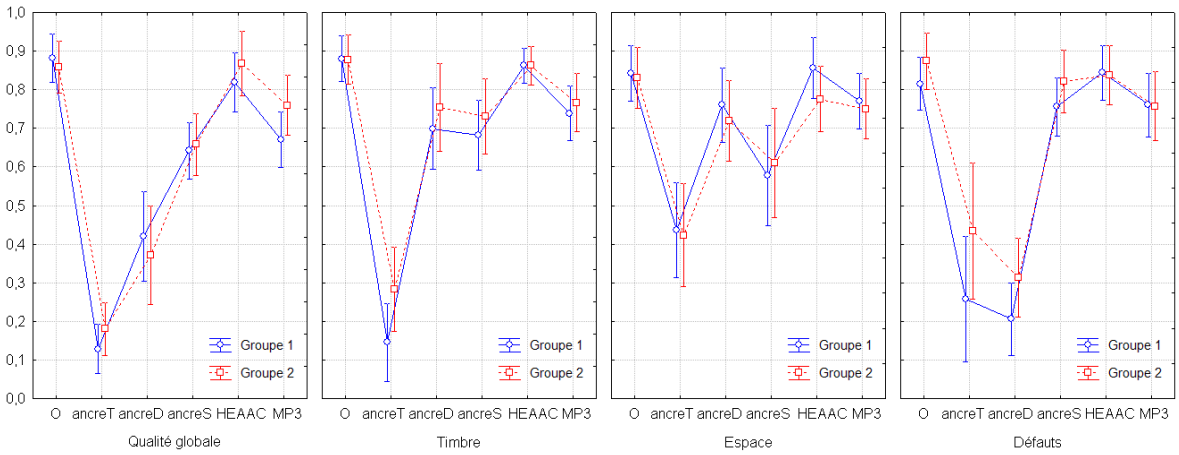


Fig. III.4 – Moyennes et intervalles de confiance à 95% obtenus par les différentes versions selon le mode de présentation des attributs *Timbre Espace* et *Défauts* successivement (Groupe 1) et simultanément (Groupe 2).

La qualité globale a été évaluée seule quel que soit le groupe d'auditeurs, le mode de présentation n'est donc pas un facteur de variabilité. Les deux groupes de sujets donnent des résultats équivalents. Pour chaque attribut évalué *Timbre, Espace* ou *Défauts*, les moyennes obtenues sont statistiquement similaires, que la présentation soit successive ou simultanée.

III.2.2 Analyse de l'évaluation selon les attributs

Le t-test montre que les deux méthodes de présentation des attributs sont statistiquement équivalentes. Il n'est donc pas nécessaire que les deux groupes d'auditeurs effectuent les tests en inversant leur mode de présentation. Les données résultantes des deux tests sont fusionnées et les analyses sont réalisées sur l'ensemble des sujets (24) en s'affranchissant désormais du mode de présentation des attributs.

Une analyse ANOVA est réalisée pour chacun des 4 attributs évalués : la qualité globale, le *Timbre*, l'*Espace* et les *Défauts*. Pour chaque attribut, les différentes versions se révèlent avoir un effet significatif sur la notation ($p<0.0001$). En revanche, les extraits présentent un effet significatif uniquement pour le *Timbre* ($F=2.6$, $p=0.032$) et pour l'*Espace* ($F=7.76$, $p<0.0001$). Un test post-hoc de Tukey montre que cette observation est due aux extraits pris individuellement et non à un type de contenu (musique, paysage, film) ou groupe d'extraits.

La figure III.5 illustre les notes moyennes et les intervalles de confiance à 95% de chaque version pour chacun des quatre attributs évalués avec les deux modes de présentation confondus.

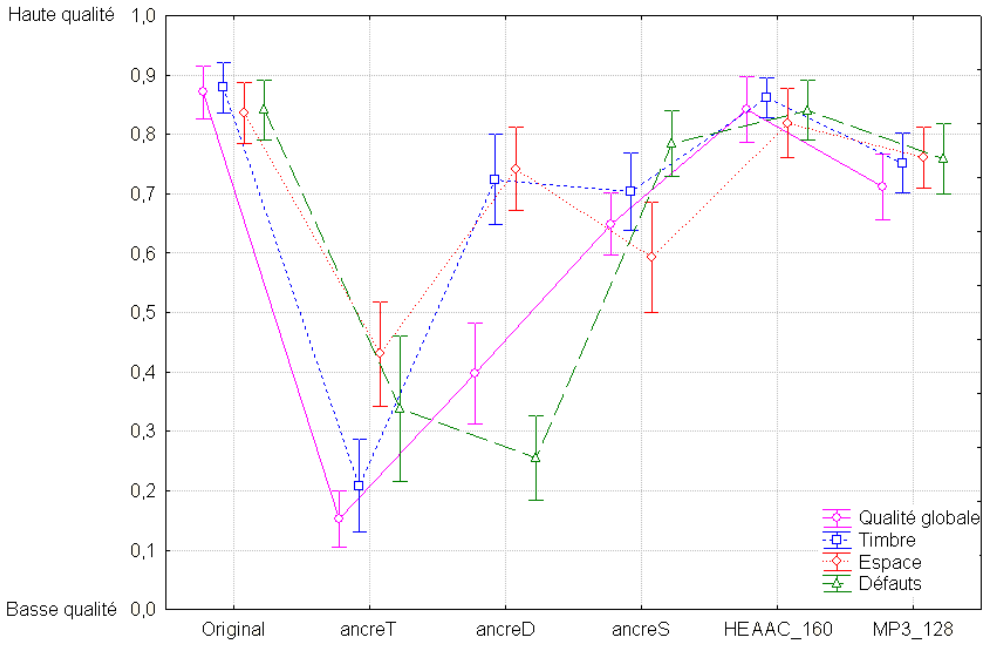


Fig. III.5 – Moyennes et intervalles de confiance à 95% des différentes versions pour chaque attribut évalué.

La version "Original", qui sert de référence cachée, est évaluée comme étant la version de plus haute qualité pour chaque attribut. Cependant, un test de Tukey révèle que le codage "HEAAC-160" est statistiquement équivalent à cette version originale pour chaque attribut (tableau III.3).

Tab. III.3 – Valeurs du post-hoc de Tukey entre la version “Original” et “HEAAC-160”.

Attribut	Qualité globale	Timbre	Espace	Défauts
Valeur du test de Tukey	0.96	0.995	0.998	1

L’absence de référence explicite montre qu’un objet évalué, ici une version dégradé de l’original, peut obtenir une moyenne équivalente voir supérieure à cette dit référence. L’auditeur étant exempté des taches d’identification et de comparaison à cette référence, il peut alors se focaliser uniquement sur l’évaluation de la qualité inter-objets. L’auditeur ne pouvait contourner la consigne de noter au moins une séquence au maximum de l’échelle. L’original n’a donc pas toujours été évalué comme étant la version de meilleure qualité et donc conforte l’hypothèse d’un test réalisé sans référence explicite pour une évaluation de qualité et non de fidélité. Une dynamique importante de l’échelle est utilisée lors de l’évaluation. Par exemple, pour la qualité globale, les moyennes vont de [0.16-0.88]. Cette constatation repose sur la présence d’ancres basse qualité et la consigne de noter une version au maximum de l’échelle et montre la pertinence des consignes et prérequis de la méthode mise en place.

L’ancrage timbral “ancreT” est différent de toutes les autres versions pour chaque attribut évalué.

Deux groupes de versions se distinguent statistiquement. Le premier contient l’“Original” et l’“HEAAC-160”. Le second dépend de l’attribut évalué et comprend le “MP3-128” et les ancrages des autres attributs. En effet, l’analyse de l’attribut *Timbre* montre que les versions “MP3-128”, “ancreD” et “ancreS” n’ont pas de différences significatives. Pour l’attribut *Espace*, “MP3-128” est similaire à “ancreD” (Valeur de Tukey : 0.997) et pour les *Défauts* “MP3-128” est équivalent à “ancreS” (VT : 0.977).

En résumé, selon l’analyse de l’attribut qui lui est associé, un ancrage est significativement différent des autres versions évaluées.

Le “MP3-128” comme l’“HEAAC-160” obtiennent des moyennes élevées et similaires selon les attributs, par exemple “MP3-128” obtient 0.72 pour la qualité globale, 0.75 pour le *Timbre*, 0.76 pour l’*Espace* et 0.76 pour les *Défauts*. Par conséquent, il est difficile d’établir des discriminations selon les attributs et donc de déterminer si un codage dégrade majoritairement certaines caractéristiques du son. Ceci peut être expliqué par la qualité haute des codages appliqués dans ce test qui ont obtenu des moyennes très élevées.

III.2.3 Le choix des ancrages

Le test inclut trois ancrages spécifiques à chaque attribut évalué. Pour l’évaluation du *Timbre*, l’ancrage “ancreT” est noté en basse qualité (0.21). Par contre, cette version est notée dans la partie inférieure de l’échelle pour les autres attributs. Le filtrage passe-bas à 3.5 kHz semble donc affecter plusieurs aspects du son dont l’*Espace* et les *Défauts* et pas uniquement le *Timbre*.

Pour l’attribut *Défauts*, l’ancrage “ancreD” est la version notée la plus basse (0.27) et par opposition il est noté en haute qualité pour les attributs *Espace* et *Timbre*. Il peut être considéré comme le bon ancrage pour cet attribut.

L'ancrage spatial "ancreS" obtient 0.6 de moyenne lors de l'évaluation de l'attribut *Espace*. Cette moyenne est supérieure à celle de l'ancrage timbral "ancreT" pour l'attribut *Espace*. Cet ancrage n'est donc pas noté en basse qualité pour l'*Espace* et ceci pour aucun des extraits évalués comme le montre la figure III.6.

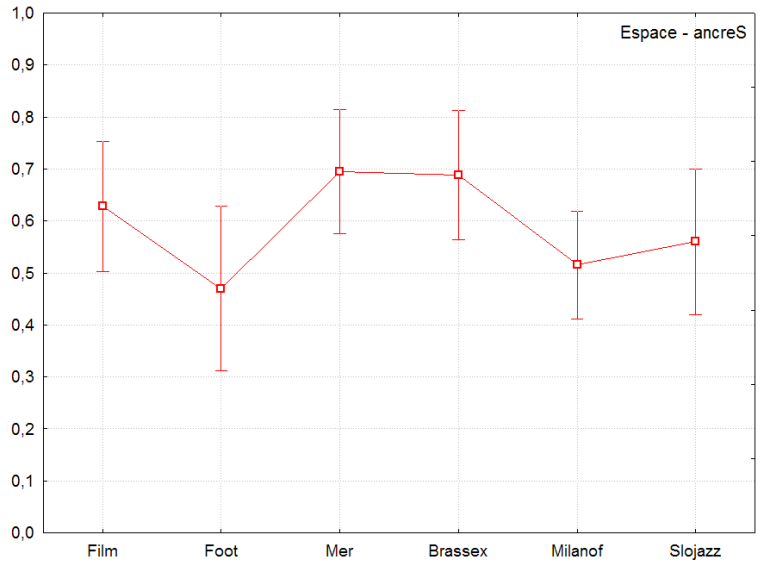


Fig. III.6 – Moyennes et intervalles de confiance à 95% de la version "ancreS" pour chaque extrait evalué lors de l'évaluation de l'attribut *Espace*.

Il a été remarqué dans d'autres études que l'ancrage spatial était noté dans la partie centrale de l'échelle de notation. Mason *et al.* (2007) avaient procédé à une réduction de la largeur de l'image sonore comme ancrage spatial. Celui-ci avait été noté à environ 56 sur 100 sur l'échelle MUSHRA. Zielinski *et al.* (2003) avait utilisé une réduction monophonique comme ancrage et il avait obtenu la note de 45 sur 100. Ceci pose la question de la dégradation spatiale. En somme, est-il possible, dans le cas d'un test sans référence, de dégrader fortement l'aspect spatial ?

III.2.4 Corrélation et régression linéaire

Le tableau III.4 référence les valeurs de corrélation des attributs. La qualité globale est fortement corrélée à chacun des trois attributs (*Timbre* : 0.87, *Espace* : 0.78, *Défauts* : 0.9). Par ailleurs les deux attributs les plus corrélés entre eux sont le timbre et l'espace (0.88).

Tab. III.4 – Valeurs de corrélation entre les quatre attributs de qualité.

Attributs	Timbre	Espace	Défauts
Qualité globale	0.87	0.78	0.90
Timbre	-	0.88	0.64
Espace	-	-	0.49

Une régression linéaire est menée dans le but de quantifier le poids de chacun sur la qualité globale. Les résultats indiquent que la qualité globale prédite et la qualité globale observée sont très proches avec une valeur de R égale à 0.985 et une valeur faible de l'erreur standard d'estimation à 0.05. La valeur de R^2 est de 0.967 et ainsi 97% de la variance de

la qualité globale peut être expliquée par les trois attributs *Timbre, Espace* et *Défauts*. La figure III.7 représente les valeurs observées et prédites de la qualité globale. Au vu de tous ces résultats le modèle montre une bonne précision.

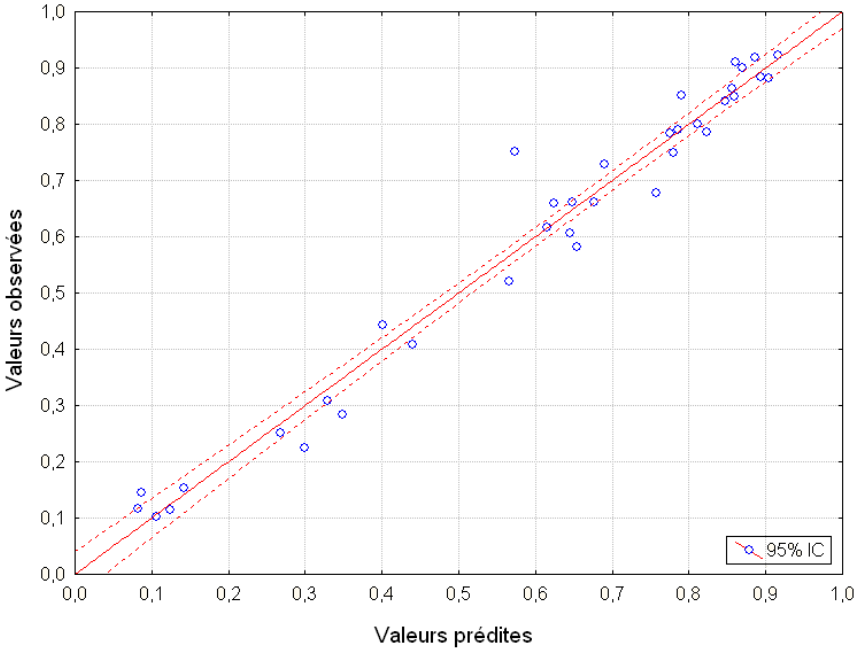


Fig. III.7 – Qualité globale, valeurs prédites vs valeurs observées.

Un des buts de cette étude est de trouver le poids de chaque attribut vis-à-vis de la qualité globale. Les coefficients de régression standardisés (β) expriment, de manières standardisées (en écarts type), l'impact exercé par une des variables indépendantes, ici le *Timbre*, l'*Espace* ou les *Défauts*, sur la variable dépendante, ici la qualité globale (Dancey *et al.*, 2007). Leurs valeurs sont de 0.25 pour l'*Espace* et le *Timbre* et de 0.61 pour les *Défauts*, qui est donc l'attribut le plus discriminant, le plus influent sur l'évaluation de la qualité globale. Les valeurs de corrélation montrent que la qualité globale est corrélée à chacun des trois attributs. Cependant, les coefficients β, montrent que la contribution sur la qualité globale du *Timbre* et de l'*Espace* est faible. Les coordonnées de l'équation de régression sont données par les coefficients non standardisés. Cette équation (III.1) permet de prédire théoriquement la note de qualité globale (QG) d'une séquence en fonction des notes obtenues pour les attributs *Timbre, Espace* et *Défauts*.

$$QG = 0.65\ \textit{Défauts} + 0.44\ \textit{Timbre} + 0.3\ \textit{Espace} - 0.32, \tag{III.1}$$

Rumsey *et al.* (2005) et Marins *et al.* (2008) ont montré que la fidélité timbrale a une influence nettement plus importante sur la qualité audio de base (BAQ) que la fidélité spatiale. La prise en compte ici d'un troisième attribut, en l'occurrence les *Défauts*, minimise le poids du *Timbre* qui reste malgré tout supérieur à l'*Espace*. Toutefois, la comparaison entre les tests précédents et l'étude actuelle est à réaliser avec précautions. Effectivement, notre test inclut uniquement deux codages et par conséquent la présence des ancrages affectent fortement les résultats de la régression linéaire. De plus, les auditeurs n'ont pas de référence explicite à laquelle se fier et donc évaluent la qualité et non la fidélité.

III.3 Conclusion

Trois attributs de qualité ont été évalués par un panel d'auditeurs à l'aide une méthode inspirée de la méthode MUSHRA.

Il a été prouvé que l'évaluation simultanée des attributs *Timbre, Espace* et *Défauts* n'engendre pas de différence en comparaison de leur évaluation successive. Cependant, elle apporte un gain de temps considérable, permettant de réduire la fatigue des auditeurs.

Cette méthode est destinée à l'évaluation de qualités dites intermédiaires. En effet, l'évaluation de codages haute qualité ne permet pas de faire de discrimination selon les attributs proposés dans ce test réalisé sans référence explicite. Il faut donc faire appel à la recommandation ITU-R BS1116 pour ces degrés de qualité.

Le choix des ancrages *Timbre* et *Défauts* a été validé. L'ancrage spatial, lui, n'a pas joué son rôle en obtenant une note moyenne élevée lors de l'évaluation de l'espace. Cette observation a été constatée dans d'autres études. Le choix de l'ancrage spatial reste donc à étudier.

La régression linéaire a révélé que l'attribut *Défauts* a plus d'influence sur la qualité globale que le *Timbre* et l'*Espace*. La présence des ancres et du nombre limité de codages, 2 seulement, impactent considérablement les résultats de cette régression.

Chapitre IV

L'application d'une méthode d'évaluation multicritère à la restitution binaurale

Suite aux premières études menées (chapitres II et III), trois attributs ont été définis et inclus dans un test d'écoute comme critères de qualité pour l'évaluation de contenus spatialisés. Il s'agit du *Timbre*, de l'*Espace* et des *Défauts*. La méthode développée est destinée à l'évaluation des sons spatialisés de manière générale. Dans ce contexte, la spatialisation peut être différemment perçue suivant le système de diffusion. De plus, l'ancrage spécifique à l'attribut *Espace* peut devoir être adapté. Dans le test d'écoute réalisé dans le chapitre III, les séquences étaient diffusées sur enceintes avec un système 5.1. Dans ce chapitre, la méthode est appliquée à des contenus binauraux restitués au casque.

IV.1 Protocole expérimental

IV.1.1 Sujets et conditions d'écoute

Dix-huit testeurs experts ont pris part à ce test.

Les écoutes se sont déroulées dans une cabine insonorisée construite spécifiquement pour la réalisation des tests d'écoute subjectifs. Les séquences sont restituées sur un casque STAX Signature SR-404 (ouvert) et son amplificateur SRM-006t.

IV.1.2 Séquences sonores

Sept extraits sonores sont choisis pour ce test et couvrent une large diversité de sons, de la musique, des extraits de films, des scènes du quotidien... Trois d'entre eux sont issus d'enregistrements binauraux natifs (micros binauraux sur tête artificielle). Pour générer les quatre autres séquences, un algorithme de synthèse binaurale est appliqué à des séquences 5.1. Ce module, nommé "VLEncoder", a été développé à Orange Labs. L'intérêt de ce procédé est de simuler un système 5.1 en positionnant les sources dans l'espace virtuel. Le but est de recréer le rendu 5.1 afin que l'auditeur ait l'impression d'être placé au centre (sweet spot) des cinq enceintes virtuelles. L'encodage binaural est réalisé à l'aide de HRTF moyennes implémentées dans le module (base de HRTF de Jean-Marie Pernaux,Pernaux (2003)). La figure IV.1 présente un exemple d'utilisation du module "VLEncoder".

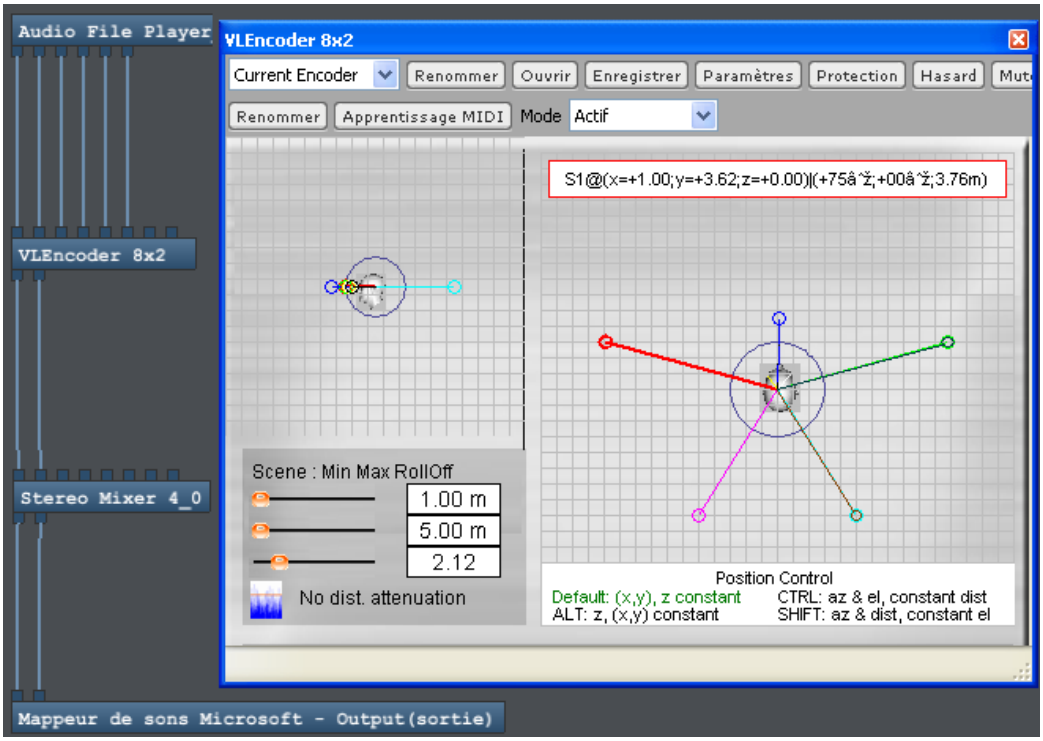


Fig. IV.1 – Exemple d'utilisation du module de synthèse binaurale "VLEncoder".

Les sept extraits sélectionnés sont décrits dans le tableau IV.1.

Tab. IV.1 – Description des extraits sonores.

Nom	*Nature*	*Description*	*Durée (s)*
Barber	Natif	Une coupe de cheveux virtuelle, des coups de ciseaux, l'utilisation de la tondeuse près des oreilles, un joueur de guitare en arrière-plan	18.8
Bombarde	Natif	Répétition de spectacle : un joueur de bombarde qui s'éloigne vers la gauche et des personnes qui préparent l'évènement en arrière plan sonore	22.5
Escalier	Natif	Deux personnes qui descendent un escalier en discutant puis qui vont dans des directions opposées	17.3
Marimba	Synthèse	Musique jouée au marimba avec effets de rotation	20.4
Milanof	Synthèse	Sons divers, enfants, tonnerre et une voix tournante	20.1
Starwars	Synthèse	Extrait de film, bataille entre vaisseaux spatiaux, tirs	18.2
Tango	Synthèse	Musique : percussion, accordéons, trompettes	18.4

Pour chacun de ces sept extraits, huit versions sont soumises à évaluation : l'original en tant que référence cachée, un ancrage spécifique à chaque attribut et quatre codages audio (tableau IV.2).

Tab. IV.2 – Description des versions évaluées.

Nom	*Description*
Original	La référence : le fichier original
ancreT	Filtrage à 3.5 kHz (ancrage timbral)
ancreD	Ajout de bruit rose et clics (ancrage défauts)
ancreS	Inversion des canaux R et L par portion + passages mono (ancrage spatial)
HEAACv2	HE-AACv2 à 40 kbits/s
AMR	AMR WB+ à 48 kbits/s
MP3	MP3 à 64 kbits/s
AAC	AAC à 32 kbits/s

Quatre codages, deux de plus que le test précédent (chapitre III), sont choisis de façon à couvrir plusieurs degrés de qualité tout en conservant le postulat d'une méthode adaptée à l'évaluation de contenus présentant des dégradations moyennes et fortes. Les codages MP3 et AAC sont conservés cependant leurs débits ont été réduits.

Le codage AMR-WB (Adaptive Multi-Rate Wideband) est un format de compression audio souvent utilisé en téléphonie mobile notamment pour le codage de la parole. Il se base sur une modélisation du système de production de la parole, la technologie ACELP (Algebraic Code Excited Linear Prediction) et utilise un détecteur d'activité vocale qui permet de transmettre uniquement les signaux montrant une activité vocale (ITU-T G.722-2, 2002). L'AMR-WB+ est une extension de l'AMR-WB (Makinen *et al.*, 2005).

HEAACv2 (High Efficiency Advanced Audio Coding) est un profil du codage AAC, complété des outils de reconstruction de bande spectrale (SBR : Spectral Band Replication) et de stéréo paramétrique (PS : Parametric Stereo) (Meltzer et Moser, 2006). Il est utilisé pour l'audio sur mobile et de plus en plus pour la radio sur internet.

Une des difficultés du montage du test réside dans le choix des ancrages spécifiques à chaque attribut. Dans l'idéal, les dégradations appliquées pour générer un ancrage doivent affecter uniquement l'attribut qui lui est associé. Par exemple, l'ancrage timbral doit entraîner la perception d'une dégradation de l'attribut *Timbre* sans altérer l'*Espace* ni les *Défauts*.

Comme dans le chapitre III, l'ancrage timbral choisi est un filtrage de la version originale avec un filtre passe-bas butterworth d'ordre 8 avec une fréquence de coupure à 3.5 kHz.

L'ancrage *Défauts* consiste à ajouter un bruit rose sur la version originale. Étant des artefacts courants, des clics sont également ajoutés. Le rapport signal sur bruit (RSB) est d'environ 30 dB. La figure IV.2 illustre la séquence de clics et le bruit rose ajoutés à la version originale pour créer l'ancrage *Défauts*.

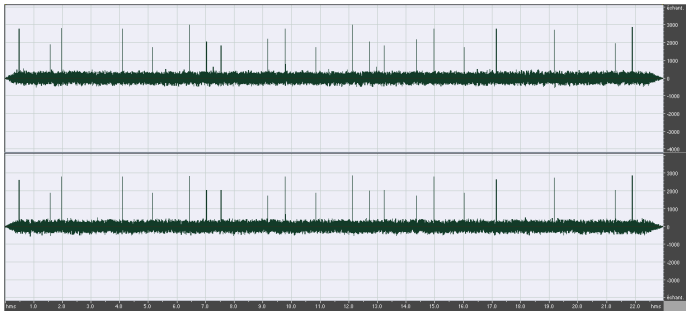


Fig. IV.2 – Séquence de clics et bruit rose ajoutés à la version originale pour créer l'ancrage *Défauts*.

Dans le test précédent, réalisé sur un système 5.1, il a été observé que l'ancrage spatial ne jouait pas son rôle. Il a obtenu une moyenne de 0.6 sur 1 lors de l'évaluation de l'attribut *Espace*. La dégradation appliquée était une inversion du canal R et Ls durant l'intégralité de l'extrait. Elle n'apportait pas suffisamment d'incohérence spatiale pour que l'ancrage soit évalué dans la partie inférieure de l'échelle de qualité proposée (échelle sans label intermédiaire avec les termes "basse" et "haute qualité" aux extrémités). D'autant plus, le test ne présente pas de référence explicite. Il est donc difficile de déceler cette altération. Pour l'ancrage spatial inclus dans ce chapitre, le choix s'est porté sur une version dynamique d'inversion des canaux. En effet, sur une courte période, le canal droit R et le canal gauche L sont échangés dans le but de créer une instabilité dans la cohérence spatiale. De plus, des passages de quelques secondes en mono sont insérés afin de générer des dégradations différentes comme une modification de localisation et d'enveloppement. La figure IV.3 décrit l'ancrage spatial.

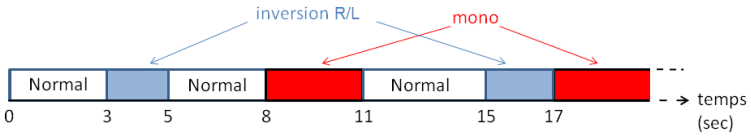


Fig. IV.3 – Description de l'ancrage spatial.

La figure IV.4 résume les dégradations appliquées aux fichiers stéréo originaux pour générer les signaux d'ancrages.

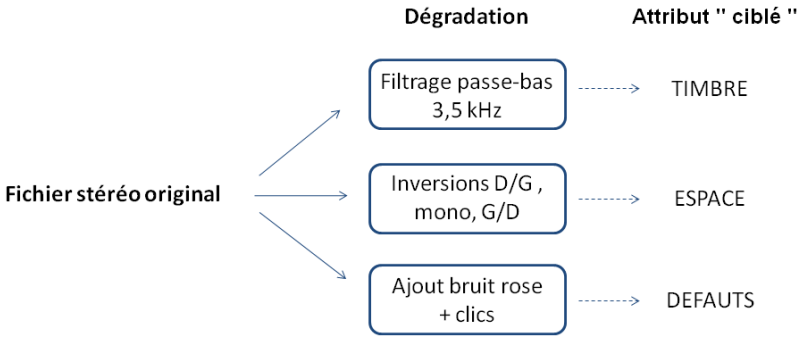


Fig. IV.4 – Description des ancrages spécifiques à chaque attribut.

IV.1.3 Déroulement du test

Le déroulement du test est similaire à celui mis en place dans le chapitre III. Le principe du test se base sur la méthode MUSHRA en s'appuyant sur une comparaison multiple simultanée des extraits sonores. Le test n'inclut pas de référence explicite pour une évaluation de qualité et non de fidélité. La consigne principale donnée aux auditeurs est de noter obligatoirement au maximum de l'échelle, l'extrait qu'ils perçoivent comme étant de plus haute qualité. Les consignes de test données aux participants sont détaillées en annexe A.3 et A.4. Le test se déroule en deux sessions conformément au chapitre III. La première consiste à évaluer la qualité globale du son, la seconde permet d'évaluer simultanément les trois attributs : *Timbre, Espace* et *Défauts*. Les interfaces de test ont été réutilisées (figures III.1 et III.3).

IV.2 Résultats

Le but de ce test est de vérifier le comportement de la méthode appliquée à un autre mode de restitution et de valider le choix des ancrages. Le test a duré en moyenne 1 heure et 51 minutes : 41 minutes pour l'évaluation de la qualité globale et 55 minutes pour l'évaluation des trois attributs. Les auditeurs étaient libres de faire des pauses. Un battement de 15 minutes était imposé entre les deux sessions.

IV.2.1 L'évaluation de la qualité globale

Les résultats de l'évaluation de la qualité globale sont présentés dans la figure IV.5. Les notes des huit versions sont obtenues par une moyenne sur l'ensemble des participants et sur la totalité des extraits proposés. Les intervalles de confiance à 95% sont également représentés.

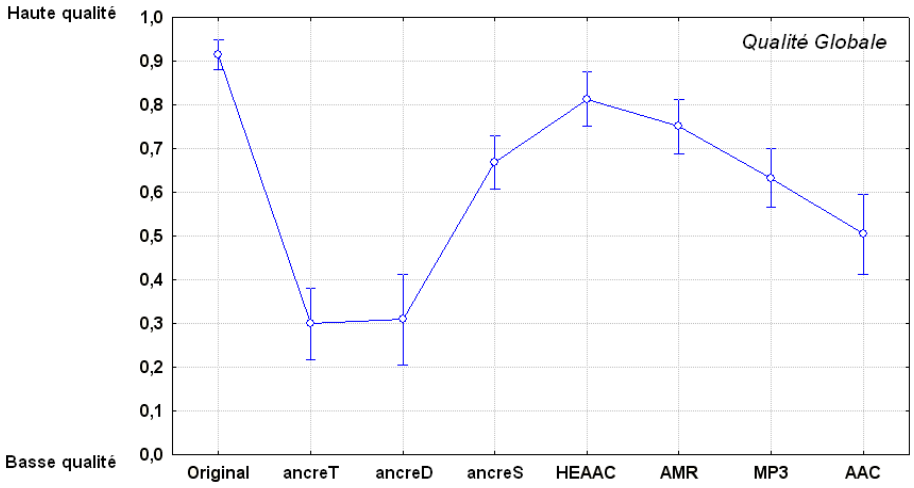


Fig. IV.5 – Moyennes et intervalles de confiance à 95% pour l'évaluation de la qualité globale.

Dans un premier temps, il est important de souligner qu'une dynamique étendue de l'échelle a été utilisée par les auditeurs pour leur notation. Ceci a été rendu possible grâce à une large plage de qualité des encodages évalués.

La meilleure moyenne est obtenue par la version originale avec 0.92 et est donc identifiée comme la version de plus haute qualité de ce test.

A l'inverse, l'ancrage timbral "ancreT" et l'ancrage *Défauts* "ancreD" obtiennent les moyennes les plus basses avec la note de 0.3. L'ancrage spatial a été noté à 0.67, une note nettement plus élevée que celles des deux autres ancres. Ceci tend à montrer que, pour l'évaluation de la qualité globale, l'aspect spatial est une dégradation secondaire. D'autre part, on peut aussi penser que l'ancrage spatial choisi est mal adapté.

Les quatre codages ont été notés dans la partie supérieure de l'échelle (> 0.5) avec une certaine hiérarchie : l'"HEAAC" obtient la note de 0.81, l'"AMR" : 0.75, le "MP3" : 0.63 et l'"AAC" : 0.50. Un post-hoc LSD (least significant difference) de Fisher permet de mettre en évidence des différences significatives entre chaque codage, excepté entre les codages "HEAAC" et "AMR" où la différence est non significative. Les valeurs du post-hoc sont indiquées dans le tableau IV.3.

Tab. IV.3 – Les valeurs du post-hoc LSD de Fisher entre les huit versions évaluées.

Moyenne *Version*	0.92 Original	0.30 ancreT	0.31 ancreD	0.67 ancreS	0.81 HEAAC	0.75 AMR	0.63 MP3	0.50 AAC
Original		0.000000	0.000000	0.000000	0.010490	0.000050	0.000000	0.000000
ancreT	-		0.789674	0.000000	0.000000	0.000000	0.000000	0.000001
ancreD	-	-		0.000000	0.000000	0.000000	0.000000	0.000002
ancreS	-	-	-		0.000340	0.039677	0.362881	0.000054
HEAAC	-	-	-	-		0.110028	0.000010	0.000000
AMR	-	-	-	-	-		0.003356	0.000000
MP3	-	-	-	-	-	-		0.001376

IV.2.2 L'évaluation des trois attributs

La seconde session du test est consacrée à l'évaluation des trois attributs *Timbre, Espace* et *Défauts*. La figure IV.6 représente les moyennes et les intervalles de confiance à 95% des versions évaluées obtenues pour chacun des attributs.

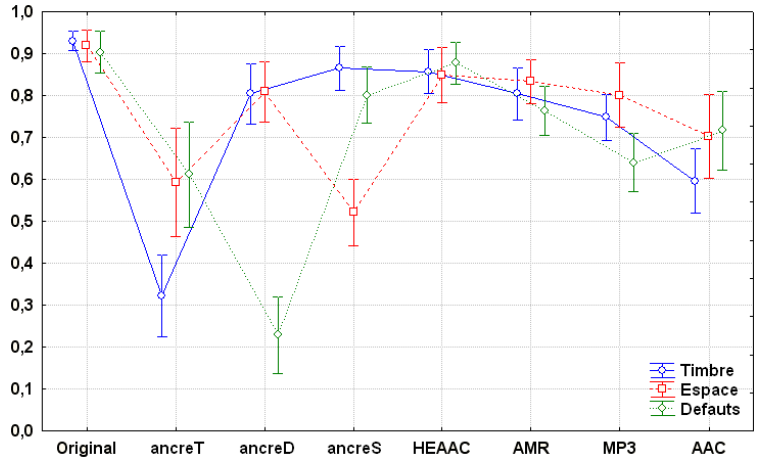


Fig. IV.6 – Moyennes et intervalles de confiance à 95% pour les 3 attributs.

Pour les trois attributs, la version originale obtient les scores les plus élevés (> 0.9).

La méthode mise en place suggère un ancrage spécifique pour chaque attribut évalué. La figure IV.6 montre que chaque ancrage est perçu comme étant la version de plus basse qualité lors de l'évaluation de l'attribut qui lui est associé. La version "ancreT" est la plus mauvaise version pour l'évaluation du *Timbre*. Pour l'évaluation de l'*Espace*, c'est l'ancrage 'ancreS' qui obtient le plus mauvais score. Pour l'attribut *Défaut*, "ancreD" obtient la plus mauvaise note. De ce fait, les testeurs semblent avoir eu une bonne compréhension des attributs.

Les deux codages "HEAAC" et "AMR" sont évalués en haute qualité et il n'y a pas de différence significative entre les trois attributs. En effet, les moyennes du *Timbre*, de *Espace* et des *Défauts* sont toutes trois similaires pour chacun de ces deux codages.

Par contre, en ce qui concerne les codages "MP3" et "AAC", des différences significatives sont constatées entre les attributs. En effet les codages semblent dégrader la qualité selon différents axes perceptifs. La note la plus basse pour l'"AAC" concerne le *Timbre* alors que la note la plus basse pour le 'MP3' concerne l'attribut *Défauts*. Le codage "AAC" dégraderait plutôt l'aspect timbral du son et le "MP3" serait à l'origine de défauts ajoutés sur les séquences. Ces observations montrent bien l'intérêt d'une méthode d'évaluation multicritères dans le but de connaître le ou les aspects du son dégradés par des systèmes de codages.

Concernant l'évaluation de l'attribut *Espace*, l'ancrage associé "ancreS" obtient la plus mauvaise note, mais est évalué à 0.52 de moyenne et se place au centre de l'échelle de qualité. Dans le test réalisé au chapitre III et dans plusieurs études incluant un ancrage spatial, celui-ci était noté au milieu de l'échelle (Mason *et al.*, 2007; Zielinski *et al.*, 2003). Pour comprendre ce phénomène, nous avons observé les notes de l'ancrage spatial "ancreS" pour chaque extrait lorsque l'évaluation portait sur l'attribut *Espace* (figure IV.7).

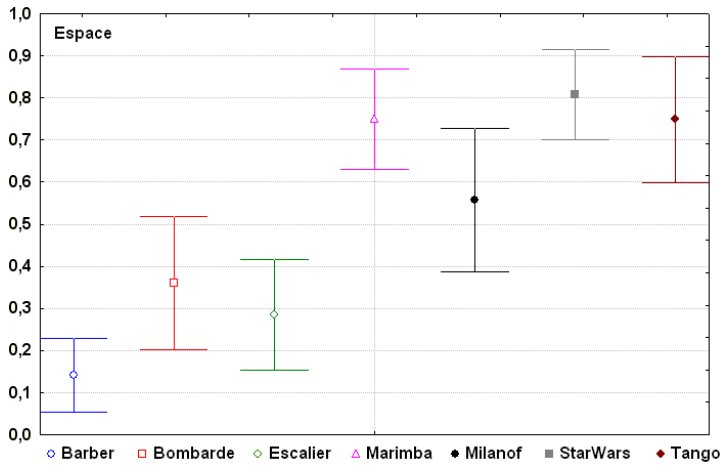


Fig. IV.7 – Moyennes et intervalles de confiance à 95% de l'ancrage spatial pour l'évaluation de l'*Espace*.

Deux groupes apparaissent clairement. Pour les trois extraits Barber, Bombarde et Escalier, l'ancrage spatial est noté en dessous de 0.4 donc en basse qualité. Le second groupe d'extraits, composé de Marimba, Milanof, Starwars et Tango, semble peu affecté

par les dégradations spatiales de l'ancrage. Cette différenciation peut venir de la méthode d'encodage binaural des extraits. En effet, les extraits Barber, Bombarde et Escalier sont issus d'enregistrements binauraux natifs tandis que les quatre autres extraits sont des séquences 5.1 binauralisées. Pour les extraits "natifs", les séquences montrent certaines similarités. En effet, la scène sonore est bien définie. Il y a une action principale et des actions secondaires en arrières plans. Leurs localisations sont précises, par exemple, une personne qui se déplace de droite à gauche en parlant. A l'inverse, les extraits initialement 5.1 contiennent plusieurs actions en parallèle (bruits de tirs et d'explosions simultanés dans plusieurs endroits de l'espace sonore). La localisation est confuse. Les dégradations spatiales appliquées étaient des inversions momentanées du canal droit et du canal gauche. La localisation plus précise des extraits "natifs" est donc plus impactée par ces inversions.

Le problème de l'ancrage spatial noté au milieu de l'échelle d'évaluation semble récurrent. Cependant, en observant la note de l'ancrage spatial en fonction des extraits, il est observé que cet ancrage est noté en basse qualité uniquement pour certains contrairement aux codages évalués (figure IV.8).

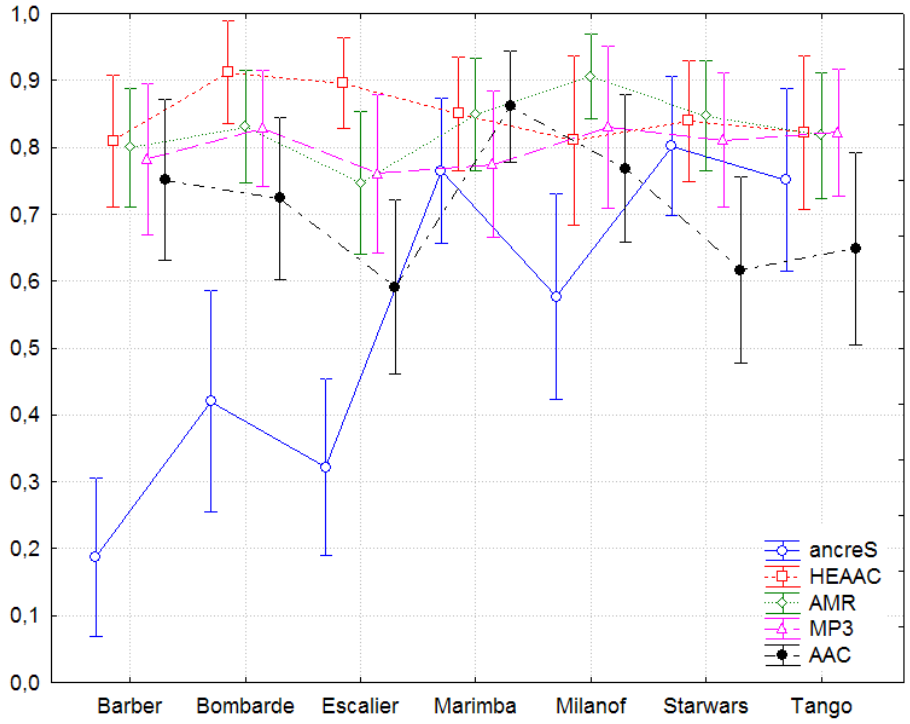


Fig. IV.8 – Moyennes et intervalles de confiance à 95% de l'ancrage spatial et des quatre codages en fonction de chaque extrait pour l'évaluation de l'*Espace*.

IV.2.3 Corrélation et régression linéaire

Le tableau IV.4 donne les valeurs de corrélation des attributs. La qualité globale est corrélée aux attributs *Timbre* (0.73) et *Défauts* (0.82) et moins à l'*Espace* (0.52). Les attributs entre eux montrent une corrélation peu élevée, en particulier l'attribut *Espace* et l'attribut *Défauts* avec une valeur de corrélation à 0.20.

Tab. IV.4 – Valeurs de corrélation entre les quatre attributs de qualité.

Attributs	Timbre	Espace	Défauts
Qualité globale	0.73	0.53	0.82
Timbre	-	0.49	0.33
Espace	-	-	0.19

Une régression linéaire est réalisée dans le but de quantifier le poids de chacun sur la qualité globale.

La valeur de R (0.968) est significative, $F = 261, p < 0.0000000$. Les trois attributs, *Timbre, Espace, Défauts* expliquent près de 94% de la variance avec $R^2 = 0.938$ et l'erreur standard d'estimation est peu importante (0.06).

Les valeurs des coefficients de régression standardisés (β) indiquent que l'attribut *Défaut* a le poids le plus important sur la qualité globale ($\beta = 0.65$). Le poids de l'attribut *Timbre* est $\beta = 0.41$ et celui de l'attribut spatial est $\beta = 0.20$. L'équation de régression est donnée par les valeurs des coefficients non standardisées.

$$QG = 0.67\ \textit{Défauts} + 0.51\ \textit{Timbre} + 0.29\ \textit{Espace} - 0.45, \tag{IV.1}$$

Cette équation permet de prédire les valeurs de qualité globale en fonction des notes obtenues pour le *Timbre*, l'*Espace* et les *Défauts* (figure IV.9).

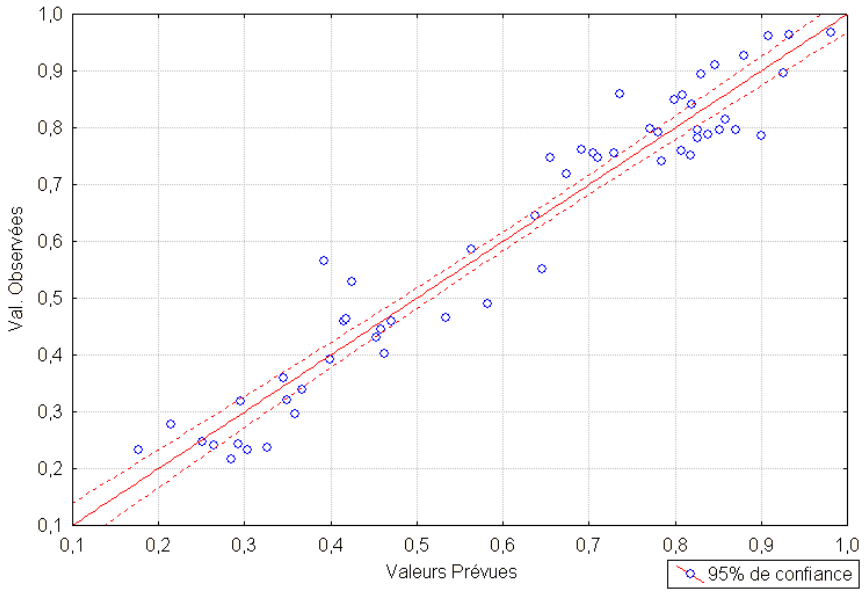


Fig. IV.9 – Prédiction de la qualité globale : valeurs observées vs valeurs prédites.

Les résultats obtenus dans le chapitre III ont montré également que l'attribut *Défauts* avait le poids le plus important sur la qualité globale. Le tableau IV.5 donne les valeurs des coefficients β obtenus lors des deux tests.

Tab. IV.5 – Comparaison des coefficients β obtenus chapitre III et chapitre IV.

Coefficient β	Timbre	Espace	Défauts
Chapitre III	0.25	0.25	0.61
Chapitre IV	0.41	0.20	0.65

Le poids de l'attribut *Timbre* est supérieur dans le test effectué au casque (0.41 contre 0.25 pour le test en 5.1). Cependant, les résultats de ces deux tests sont difficilement comparables : le système de restitution est différent, l'ancrage spatial est adapté et deux codages supplémentaires sont évalués minimisant l'influence des ancres dans la régression. Chacune de ces différences est une explication potentielle.

Les ancrages sont des références basse qualité utilisés pour permettre, entre autres, une notation sur toute la dynamique de l'échelle mais ils ne sont pas les objets de l'évaluation. Leur présence dans la régression linéaire est discutable. En effet, le poids d'un attribut est dépendant de l'évaluation de son ancrage, et, il s'avère que l'ancrage spatial ne joue pas son rôle. Nous allons nous intéresser aux quatre codages inclus dans le test et identifier les axes de qualité qui influencent leur qualité globale. Même si, d'un point de vue protocolaire/statistique, les ancrages doivent être pris en compte, une régression linéaire est réalisée sur ces quatre codages uniquement. Le tableau IV.6 donne les valeurs de corrélations entre les quatre attributs en tenant compte des quatre codages (sans ancres et sans la référence cachée).

Tab. IV.6 – Valeurs de corrélation entre les quatre attributs de qualité calculées uniquement sur les quatre codages.

Attributs	Timbre	Espace	Défauts
Qualité globale	0.89	0.73	0.81
Timbre	-	0.86	0.61
Espace	-	-	0.42

Les trois attributs *Timbre, Espace* et *Défauts* sont corrélés à la qualité globale et sont également corrélés entre eux notamment le *Timbre* et l'*Espace* (0.86).

Les valeurs des coefficients de régression standardisés (β) indiquent que l'attribut *Timbre* a le poids le plus important sur la qualité globale ($\beta = 0.59$). Le poids de l'attribut *Défauts* est $\beta = 0.44$ et celui de l'attribut spatial est $\beta = 0.03$. En prenant en compte les quatre codages uniquement (sans l'original et sans les ancrages), l'attribut *Espace* est non significatif ($p > 0.5$) et seuls le *Timbre* et les *Défauts* expliquent la variabilité de la qualité globale.

L'influence des ancrages sur les valeurs des poids est importante notamment pour l'attribut *Espace*. En restitution sonore, la dimension spatiale est peu exploitée de par un manque de contenus dédiés et des dispositifs matériels complexes. Les auditeurs peuvent éprouver des difficultés à évaluer cet aspect du son à cause d'un manque d'expérience pour ce type d'écoute. Il est également possible que les systèmes de codage affectent peu cette dimension et dégradent d'autres caractéristiques comme le *Timbre* et/ou les *Défauts*.

IV.3 Conclusion

La méthode d'évaluation de qualité incluant la notation des attributs *Timbre, Espace* et *Défauts* simultanément a été appliquée à des contenus binauraux restitués au casque.

Les quatre systèmes de codages testés sont statistiquement différents pour l'évaluation de la qualité globale, ce qui montre une certaine diversité dans les degrés de qualité proposés. L'évaluation suivant plusieurs critères d'évaluation montre son intérêt particulièrement sur les versions de qualité intermédiaire. En effet, un codage peut dégrader spécifiquement un aspect du son, par exemple, le MP3 affecte plus fortement l'attribut *Défauts* que le *Timbre* ou l'*Espace*. A l'inverse, l'AAC dégrade davantage le *Timbre* que les deux autres attributs.

L'ancrage spatial inclus dans ce test a joué son rôle pour la moitié des extraits seulement, où il a été évalué en basse qualité pour l'attribut *Espace*. Pour la seconde moitié, il a obtenu des moyennes élevées. Le choix des contenus peut donc être une source de variabilité important lors d'un test d'écoute. L'ancrage développé dans ce chapitre a apporté une certaine amélioration à propos de son évaluation en basse qualité pour l'attribut *Espace*. Cependant, des investigations sont à mener pour que celui-ci fonctionne sur l'ensemble des extraits présentés.

La régression linéaire a confirmé que l'attribut *Défauts* a un impact plus important sur la qualité globale que les attributs *Timbre* et *Espace*. Cependant, une régression linéaire sur les quatre codages, sans prendre en compte les ancrages et l'original, montrent une prédominance du *Timbre* suivi des *Défauts* et qu'ils expliquent à eux deux la variabilité de la qualité globale.

Chapitre V

La conception et le choix de l'ancrage spatial

Une méthode d'évaluation subjective des sons spatialisés a été proposée dans le chapitre III sur un système 5.1 et testée dans le chapitre IV sur des contenus binauraux avec une restitution au casque. Cette méthode propose une évaluation sur trois critères, dont l'attribut *Espace*, pour compléter l'évaluation de la qualité globale. Un ancrage spécifique à cet attribut a été inclus dans les tests précédents comme une référence implicite basse qualité. Or, il a obtenu des notes moyennes élevées. Ce chapitre est consacré au choix de l'ancrage spatial à intégrer dans la méthodologie de test. Un ancrage doit, entre autres, permettre une reproductibilité d'un test audio. Il sert de référence basse qualité à l'auditeur. Cependant, la différence de qualité avec les objets évalués doit rester cohérente. Une qualité d'ancrage trop basse par rapport à celle des objets évalués réduirait la dynamique de notation entre ces objets. Si la qualité de l'ancrage est trop élevée, il pourrait alors obtenir une moyenne supérieure à un ou plusieurs objets évalués et par conséquent ne jouerait pas son rôle. De plus, cet ancrage doit être évalué en basse qualité pour la majorité des extraits proposés dans un test.

V.1 Propositions d'ancrages spatiaux

L'ancrage spatial recherché pour cette méthode de test doit répondre aux critères suivants :
- être évalué en basse qualité pour l'attribut *Espace*.
- obtenir des moyennes proches indépendamment des extraits proposés.
- affecter uniquement l'attribut *Espace* et non le *Timbre* et les *Défauts*.
- être de qualité inférieure mais cohérente avec celle des objets du test.

Les ancrages proposés dans les chapitres III et IV ne répondent pas aux critères cités ci-dessus en obtenant notamment des moyennes élevées. La figure V.1 représente les moyennes de l'ancrage spatial "ancreS" pour chacun des attributs évalués dans les tests décrits dans les chapitres III (système 5.1) et IV (technologie binaurale).

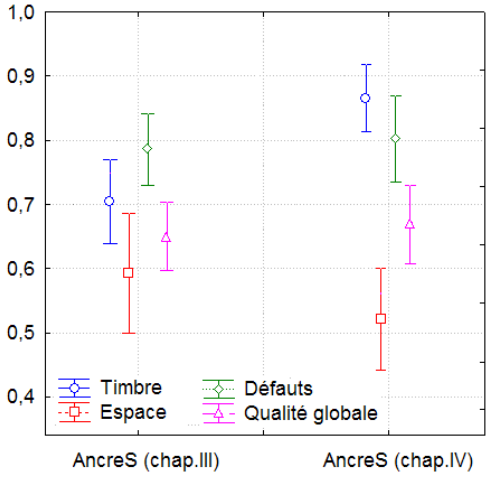


Fig. V.1 – Moyennes et intervalles de confiance obtenus par "AncreS" lors de l'évaluation de l'attribut *Espace* dans les chapitres III et IV.

L'ancrage proposé dans le chapitre III consistait en une inversion du canal avant droit R et du canal arrière gauche Ls sur toute la durée des extraits. La moyenne obtenue pour cet ancrage était élevée (0.59) pour l'attribut *Espace* et proche des moyennes des autres attributs (figure V.1). Ce procédé dégradait tout autant le *Timbre* que l'*Espace* ou les *Défauts* alors qu'il aurait du affecter principalement la perception de l'*Espace*.

Dans le chapitre IV, pour la restitution au casque, une inversion par courte période (deux secondes) du canal droit R et du canal gauche L ainsi que des périodes en mono ont été appliquées. Cet ancrage a obtenu une moyenne de 0.52, trop élevée, mais cette moyenne se différencie statistiquement de celles des attributs *Timbre* et des *Défauts* (figure V.1). Concrètement, l'inversion momentanée a davantage dégradé la qualité de l'*Espace* que les autres axes de qualité évalués mais pas suffisamment. De plus, "ancreS" a obtenu une note inférieure à 0.38 uniquement pour la moitié des extraits. La notation de cet ancrage était donc dépendante du contenu des extraits évalués.

Le but du chapitre V est de définir un ancrage basse qualité, qui dégrade essentiellement et de manière prononcée l'aspect spatial du son et cela pour tous les extraits sonores. Cinq ancrages sont proposés et soumis à validation lors d'un test d'écoute. La restitution lors du test est faite au casque dans le but d'une évaluation de contenus binauraux.

Le premier ancrage inclus dans le test est l'ancrage réalisé dans le chapitre IV afin de comparer le comportement de celui-ci aux autres possibilités d'ancrages.

Certaines études ont inclus un ancrage spatial dans leur test d'écoute. Zielinski *et al.* (2003) ont utilisé une réduction monophonique et Mason *et al.* (2007) et Marins *et al.* (2008) ont procédé à une réduction de l'image sonore. Ces deux ancrages ont été évalués dans la partie centrale de l'échelle de notation qu'ils proposaient dans leurs études. Afin de vérifier le comportement de ces deux types d'ancrage dans cette méthodologie, ils ont été inclus au corpus de stimuli.

Une version monophonique des signaux originaux a été réalisée en sommant le canal droit R et le canal gauche L. L' et R' correspondent respectivement au canal gauche et au

canal droit de l'ancrage monophonique.

$$L' = \frac{L+R}{2} \text{ et } R' = \frac{L+R}{2}$$

La réduction de l'image sonore stéréophonique consiste à ajouter au canal droit 25% du canal gauche et inversement (figure V.2).

$$L' = (1-\alpha) * L + \alpha * R \text{ et } R' = (1-\beta) * R + \beta * L \text{ ici } \alpha = \beta = 0.25$$

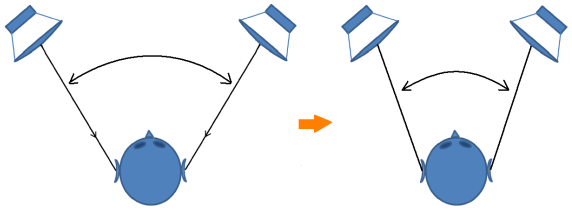


Fig. V.2 – Perception de la réduction de l'image sonore.

Le logiciel Adobe Audition© implémente un module nommé *Effet panoramique auto binaural* permettant de générer un mouvement circulaire de la gauche vers la droite sur les séquences. Le canal droit ou gauche est retardé pour donner l'impression que les sons arrivent au niveau des oreilles à des instants différents. Cet outil a été utilisé pour la conception d'une ancre spatiale. La fréquence de déplacement du son d'un canal à l'autre appliquée est de 0.175 Hz, ce qui équivaut à un mouvement temporel de 5.7 secondes (figure V.3).

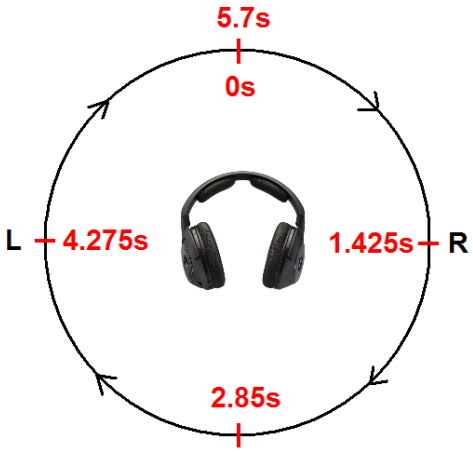


Fig. V.3 – Mouvement circulaire du son de la gauche vers la droite avec les correspondances temporelles.

L'inconvénient de l'ancrage expliqué dans le paragraphe ci-dessus est le manque d'information sur le fonctionnement interne de ce module. En conséquence, un effet similaire a été conçu et appliqué sur les extraits sonores : le signal monophonique noté M oscille entre le canal droit et le canal gauche en suivant une fonction sinus représentée par les coefficients α et β avec une période de 5 secondes. L' et R' sont respectivement le canal gauche et le canal droit de cet ancrage.

$L' = \frac{L}{V} + M * \alpha$ et $R' = \frac{R}{V} + M * \beta$

$M = \frac{L+R}{2}$ (Réduction monophonique) et V = 2 (Paramètre du volume initial du canal originel)

$\alpha = sin(2 * \pi * A * t)^2$ et $\beta = sin(2 * \pi * A * t + \frac{\pi}{2})^2$, ici A = 0.1.

Les valeurs des paramètres V et A sont choisis à l'écoute dans le but d'imiter le rendu obtenu avec le module d'Adobe Audition©.

V.2 Protocole expérimental

V.2.1 Conditions d'écoute et sujets

Vingt-deux auditeurs experts ont pris part à ce test dans les mêmes conditions que le test réalisé au chapitre IV dans une cabine insonorisée dédiée et avec une restitution au casque.

V.2.2 Stimuli

Pour pouvoir comparer les résultats avec le test réalisé dans le chapitre IV, les sept mêmes extraits ont été utilisés (Chap.IV tab. IV.1). On rappelle que trois des extraits sont des enregistrements binauraux natifs, quatre sont des séquences 5.1 binauralisées avec un module de synthèse. Trois extraits ont été rajoutés. Ces extraits, nommés "Alanis", "Dion" et "Money", sont des enregistrements stéréo musicaux présentant une spatialisation intéressante. Au total, dix extraits ont été présentés aux auditeurs (tableau V.1).

Tab. V.1 – Description des extraits sonores.

Nom	*Nature*	*Description*	*Durée (s)*
Barber	Natif	Une coupe de cheveux virtuelle, des coups de ciseaux, l'utilisation de la tondeuse près des oreilles, un joueur de guitare en arrière-plan	18.8
Bombarde	Natif	Répétition de spectacle : un joueur de bombarde qui s'éloigne vers la gauche et des personnes qui préparent l'évènement en arrière plan sonore	22.5
Escalier	Natif	Deux personnes qui descendent un escalier en discutant puis qui vont dans des directions opposées	17.3
Marimba	Synthèse	Musique jouée au marimba avec effets de rotation	20.4
Milanof	Synthèse	Sons divers, enfants, tonnerre et une voix tournante	20.1
Starwars	Synthèse	Extrait de film, bataille entre vaisseaux spatiaux, tirs	18.2
Tango	Synthèse	Musique : percussion, accordéons, trompettes	18.4
Alanis	Stéréo	Extrait de *Head over feet* interprété par Alanis Morissette Voix au centre, guitares à droite et à gauche, batterie, clavier	16
Dion	Stéréo	Introduction musicale de *Je crois toi* de Céline Dion Guitares, violoncelles, synthétiseur	15.6
Money	Stéréo	Introduction musicale de *Money* du groupe Pink Floyd Basse, batterie, guitare électrique, machine à sous et machine à écrire en alternance droite/gauche	15.8

Huit versions pour chaque extrait ont été soumises à évaluation (tableau V.2) dont cinq propositions d'ancrages (paragraphe V.1). La version originale ainsi que deux codages, HEAAC à 24 kbits/s et AAC à 32 kbits/s (chapitre IV), ont été ajoutés pour comparer les résultats des ancrages potentiels en conditions réelles de test c'est à dire en comparaison directe avec des objets évalués, ici des systèmes de codages. La dynamique de notation aurait été différente si les ancrages avaient été comparés entre eux uniquement. De plus, un des critères recommandés pour la validité de l'ancrage est qu'il obtienne une note inférieure mais cohérente à celles des objets évalués, en l'occurrence des systèmes de codages.

Tab. V.2 – Description des versions évaluées.

Original	La référence : le fichier original
InvR/L	Inversion momentanée R/L + mono (ancrage chapitre IV)
Adobe	Module de panning existant (adobe audition)
ModuleFT	Module de panning recréé
Reduc	Réduction de l'image stéréophonique
Mono	Réduction monophonique
HEAACv2	HE-AACv2 à 24 kbits/s
AAC	AAC à 32 kbits/s

V.2.3 Déroulement du test

Le test s'est déroulé en deux parties, une phase de familiarisation et la phase de test. Il a duré en moyenne 61 minutes (phase de familiarisation et test). L'unique attribut évalué est l'*Espace* qui, par la définition établie au chapitre II, fait référence à l'impression spatiale relative aux caractéristiques spatiales : la profondeur, la largeur, la localisation, la distribution spatiale, la réverbération, la spatialisation, la distance, l'enveloppement, l'immersion.

Les deux extraits sonores utilisés durant la phase d'entraînement sont les extraits "Dion" et "Escalier". Le but de cette étape est de faire découvrir aux auditeurs les différentes dégradations spatiales apportées aux séquences. Les huit autres extraits ont été utilisés lors de la phase de test. Pour les deux parties du test d'écoute, le procédé d'évaluation reste identique aux tests réalisés précédemment. L'échelle est une échelle continue de qualité avec pour extrémités, les étiquettes : basse qualité et haute qualité. La version originale est incluse comme une référence cachée, les auditeurs n'en sont pas informés. Ils doivent noter au maximum de l'échelle la version qu'ils jugent de plus haute qualité. Les huit versions d'un extrait ont été évaluées simultanément sur une même interface (figure V.4).

V.3 Résultats

Les résultats issus de la phase de familiarisation n'ont pas été pris en compte dans les analyses présentées dans cette partie. Les analyses ont été réalisées sur soixante-quatre séquences, huit extraits, chacun déclinés en huit versions à évaluer.

La figure V.5 représente les moyennes et les intervalles de confiance à 95% de chaque version.

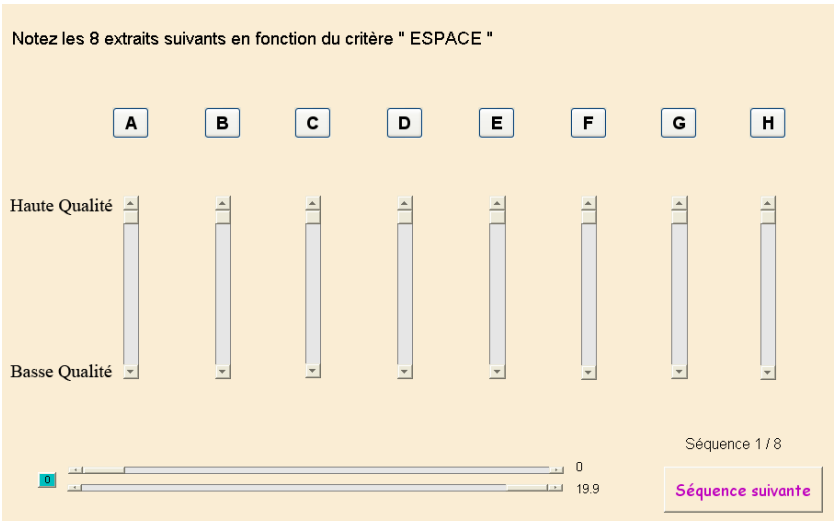


Fig. V.4 – Interface de test utilisée lors de l'évaluation de l'attribut *Espace*.

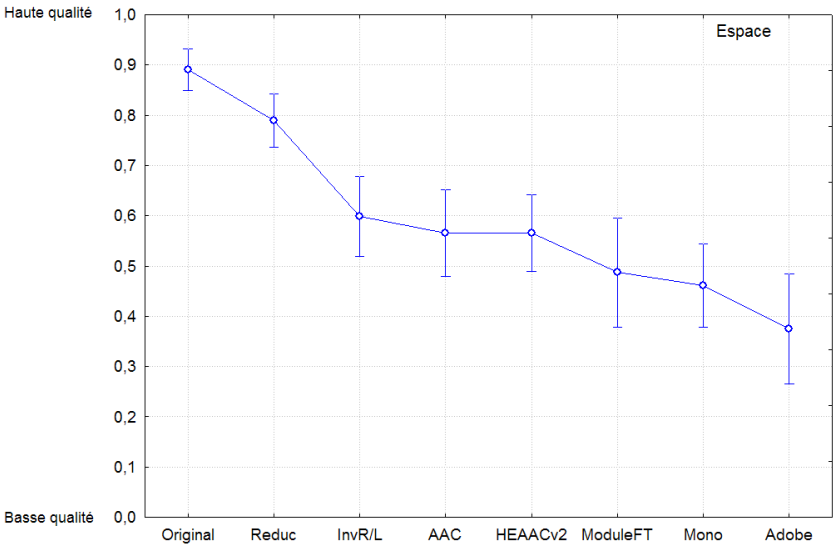


Fig. V.5 – Moyennes et intervalles de confiance à 95% pour chaque version lors de l'évaluation de l'attribut *Espace*.

"Original" est la version la mieux notée du test avec 0.89 de moyenne.

L'ancrage "Reduc" obtient une moyenne de 0.79 (haute qualité). Il ne convient donc pas comme ancrage.

La version "InvR/L" correspond à l'ancrage spatial utilisé dans le chapitre IV. Les notes de cet ancrage, évalué dans ce chapitre, sont équivalentes à celles obtenues dans le chapitre IV en comparant les moyennes sur les six extraits similaires aux deux tests (figure V.6). Concernant l'attribut *Espace*, les moyennes entre les deux tests pour chaque extrait sont statistiquement similaires d'après les valeurs du post-hoc de Tukey (Valeur = 1). Ceci montre une parfaite reproductibilité du test.

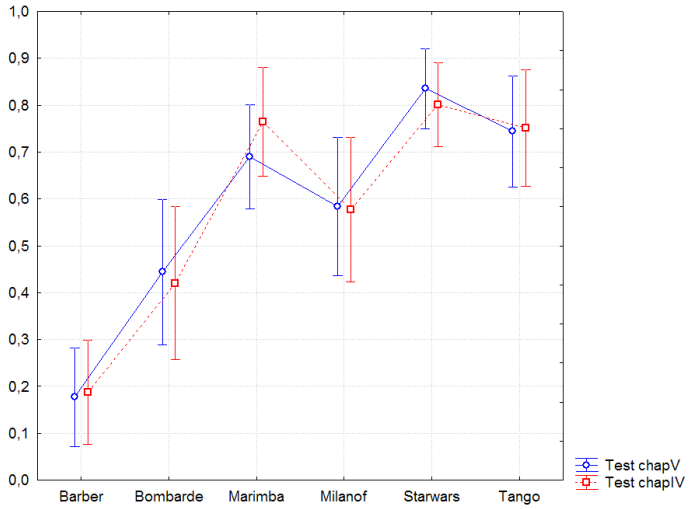


Fig. V.6 – Moyennes et intervalles de confiance à 95% de la version “InvR/L” pour les six extraits évalués à la fois dans le chapitre IV et dans le chapitre en cours.

Les inconvénients de la version “InvR/L” mis en évidence dans le chapitre IV sont à nouveau observés. Il avait été remarqué que la notation différait suivant l’extrait évalué. En effet, trois extraits sur sept ont été perçus en basse qualité, les quatre autres ont obtenues des moyennes élevées. Les dégradations spatiales ont été détectées uniquement sur la moitié des extraits ; ceci montre donc l’influence du contenu sur la notation. La figure V.7 représente les moyennes et intervalles de confiance de la version “InvR/L” en fonction de chacun des huit extraits évalués. Les extraits “Barber” et “Bombarde ” obtiennent des notes bien inférieures à celles des extraits “Starwars” et “Tango” (figure V.7). Les résultats confirment la conclusion du chapitre IV à savoir que l’ancrage “invR/L” est inapproprié car il est dépendant du contenu.

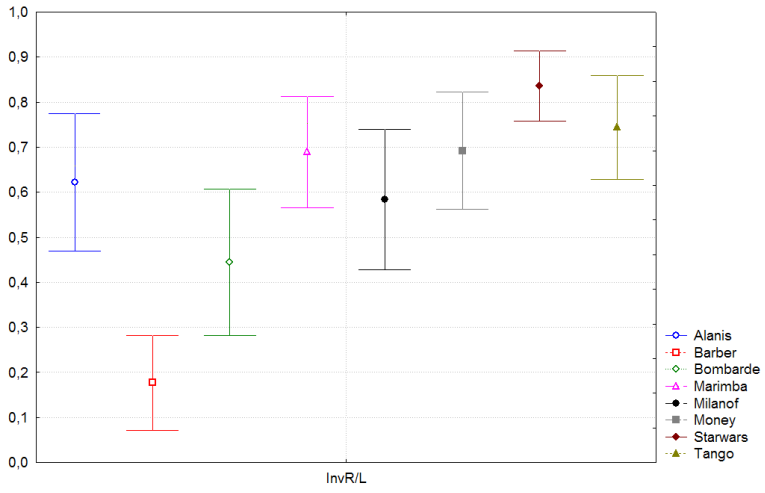


Fig. V.7 – Moyennes et intervalles de confiance à 95% de la version “InvR/L” pour les huit extraits évalués.

Les codages “AAC-32” et “HEAACv2-24” obtiennent tous deux une moyenne de 0.57. Le but, dans ce chapitre, est de définir un ancrage basse qualité à inclure dans un test

d'écoute pour l'évaluation de l'attribut *Espace*. Les codages permettent de comparer les ancrages évalués dans des conditions réelles de test. L'ancrage recherché doit obtenir une note inférieure à celles des codeurs. Trois des ancrages proposés ont une moyenne inférieure à 0.57 : "ModuleFT" avec 0.49, "Mono" avec 0.46 et "Adobe" avec 0.38 (figure V.5).

Les ancrages "ModuleFT" et "Adobe" sont notés de manière quasi égale en fonction des extraits et donc conviennent pour tout type de contenu. La dispersion des extraits sur l'échelle est plus importante pour l'ancrage "Mono" que pour "ModuleFT" et "Adobe" (figure V.8).

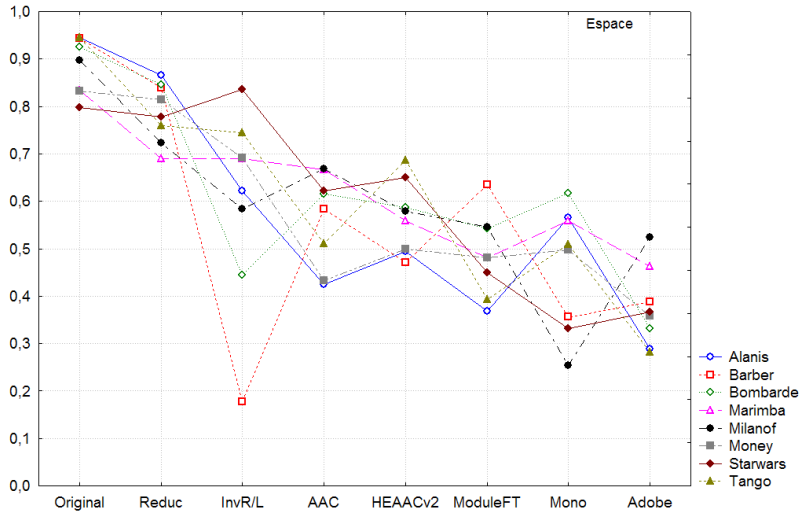


Fig. V.8 – Moyennes et intervalles de confiance à 95% pour chaque version et chaque extrait lors de l'évaluation de l'attribut *Espace*.

Les ancrages "ModuleFT", "Mono" et "Adobe" sont considérés comme statistiquement équivalent d'après le résultat d'un post-hoc de Tukey. Un post-hoc LSD de Fisher permet de dissocier statistiquement l'ancrage "Adobe" des ancrages "ModuleFT" et "Mono".

Une remarque est à souligner : les extraits qui ont obtenu la note la plus basse pour l'ancrage "Adobe" par exemple "Tango", "Alanis", "Bombarde" sont les plus hauts pour l'ancrage "Mono" et inversement. La nature dynamique des mouvements des sources dans l'espace liés aux extraits évalués n'explique pas cette observation.

V.4 Conclusion

Cinq possibilités d'ancrages spatiaux basse qualité ont été évaluées et comparées afin de déterminer l'ancrage qui répond au mieux aux critères requis dans le but d'être inclus dans la méthodologie de test : cet ancrage doit être évalué en basse qualité et ce, pour chaque extrait présenté. Les ancrages nommés "Adobe" et "ModuleFT" obtiennent les moyennes les plus basses et une dispersion faible en fonction des extraits. Ils apportent de réelles modifications spatiales notamment une incohérence dans l'espace dû à un mouvement rotatif. Ce type de dégradations répond donc au mieux aux critères exigés.

Chapitre VI

L'intégration de l'ancrage spatial

Le test réalisé dans le chapitre V a permis de mettre en évidence les dégradations qui altèrent fortement et principalement la perception spatiale des extraits sonores. Il s'agit de créer un mouvement circulaire entre l'oreille gauche et l'oreille droite. Cette dégradation va donc être appliquée et utilisée comme ancrage spatial dans la méthodologie de test mise en place dans les chapitres III et IV. Le choix de cet ancrage spécifique à l'attribut *Espace* va être validé par le biais d'un test d'écoute consacré à l'évaluation de qualité de contenus audio binauraux.

VI.1 L'ancrage spatial inclus dans le test

Le test décrit dans le chapitre V a montré que, pour chaque extrait, l'ancrage "Adobe" obtient la note la plus basse lors de l'évaluation de l'attribut *Espace*. L'inconvénient de cet ancrage est l'absence d'information sur les paramètres utilisés par ce module. La version "ModuleFT" a été conçue dans le but de reproduire les dégradations engendrées par la version "Adobe". Sa moyenne est plus élevée que celle de la version "Adobe" mais, selon un post-hoc de Tukey, les deux versions ne présentent pas de différences significatives. De ce fait, des modifications ont été apportées à la version "ModuleFT" pour être plus fidèle à la version "Adobe". La sensation produite par le module "Adobe" est un mouvement oscillant de gauche à droite avec une impression d'arrêt à droite et à gauche. Pour imiter cela, une fonction trapèze est appliquée sur toute la longueur du signal au lieu de la fonction sinus (Chapitre V.1) et la période du mouvement est de 6.54 secondes (figure VI.1).

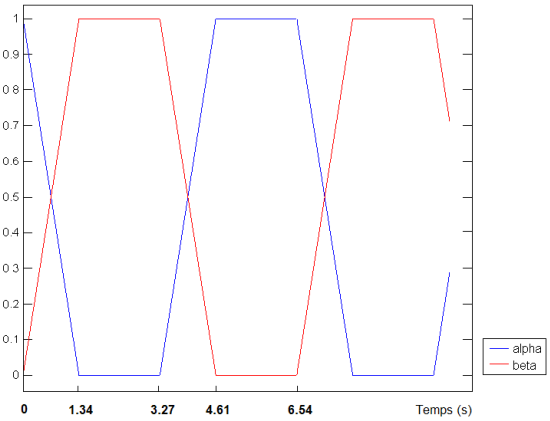


Fig. VI.1 – Tracé des coefficients α et β.

L' et R' sont respectivement le canal gauche et le canal droit de l'ancrage spatial utilisé dans ce test.

$L' = \frac{L-R}{V} + M * \alpha$ et $R' = \frac{R-L}{V} + M * \beta$

$M = \frac{L+R}{2}$ (Réduction monophonique) et V = 7

$$\alpha(t) = \begin{cases} (1 - \delta * a * t) \ si \ t \in [(0, p)] \\ 0 \ si \ t \in [p, p+n] \\ (\delta * a * t) \ si \ t \in [(p+n, 2p+n)] \\ 1 \ si \ t \in [2p+n, 2(p+n)]] \end{cases} \quad et \quad \beta(t) = 1 - \alpha(t)$$

$avec \ \delta = 0.0000001, \quad a = 170, \quad n = 1.93, \quad p = 1.34$

Les valeurs des paramètres δ, a, n, p et V sont choisis de manière perceptive afin que l'effet de rotation soit similaire à celui du module "adobe".

VI.2 Protocole expérimental

VI.2.1 Conditions d'écoute et panel d'écoute

Les conditions d'écoute sont identiques à celles des tests réalisés dans les chapitres IV et V. Les séquences sont diffusées sur un casque STAX dans une cabine insonorisée. Vingt sujets "experts" participent à ce test.

VI.2.2 Stimuli

Dix extraits sont présentés aux auditeurs (tableau V.1) :
- les sept extraits évalués lors du test d'écoute réalisé dans le chapitre IV contenant des extraits binauraux natifs et des extraits 5.1 binauralisés.
- les trois extraits stéréo "Alanis","Dion" et "Money" utilisés dans le chapitre V.

Pour chaque extrait, huit versions sont présentées (tableau VI.1) : un ancrage spécifique à chaque attribut évalué (*Timbre*, *Espace* et *Défauts*), la version originale et quatre codages audio identiques à ceux évalués dans le chapitre IV dans le but de comparer les résultats des deux tests.

Tab. VI.1 – Description des versions évaluées.

Nom	*Description*
Original	La référence : le fichier original
ancreT	Filtrage à 3.5 kHz (ancrage timbral)
ancreD	Ajout de bruit rose et clics (ancrage défauts)
ancreS	Mouvement oscillant entre le canal droit et gauche (ancrage spatial)
HEAACv2	HE-AACv2 à 40 kbits/s
AMR	AMR WB+ à 48 kbits/s
MP3	MP3 à 64 kbits/s
AAC	AAC à 32 kbits/s

VI.2.3 Déroulement du test

La première phase du test consiste à évaluer la qualité globale du son (figure VI.2).

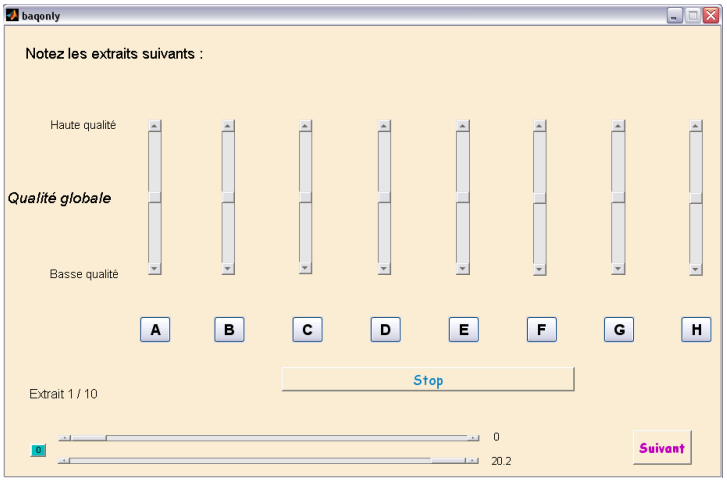


Fig. VI.2 – Interface de test pour l'évaluation de la qualité globale.

Dans la seconde partie du test, il est demandé aux auditeurs d'évaluer les séquences suivant les trois attributs de qualité définis dans le chapitre II : *Timbre*, *Espace* et *Défauts*. Les trois attributs sont présentés sur une même interface (figure VI.3).

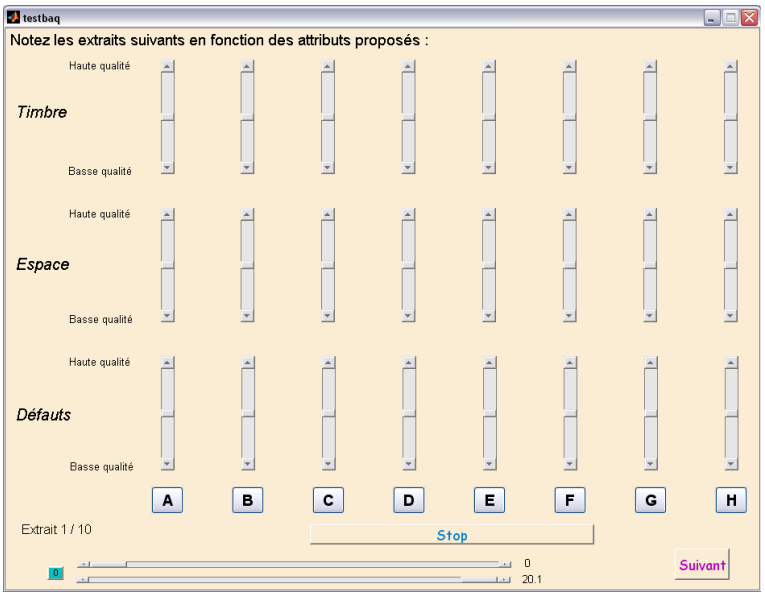


Fig. VI.3 – Interface de test pour l'évaluation des trois attributs de qualité : *Timbre, Espace, Défauts.*

La version originale n'est pas définie comme une référence explicite pour les auditeurs cependant elle fait office de référence cachée haute qualité. L'échelle de notation utilisée est une échelle de qualité avec pour extrémités les labels "Basse qualité" et "Haute qualité". Les consignes de test, données aux auditeurs, sont détaillées dans l'annexe A.3 et A.4.

VI.3 Résultats

VI.3.1 Qualité globale

Les résultats de l'évaluation de la qualité globale sont présentés sur la figure VI.4.

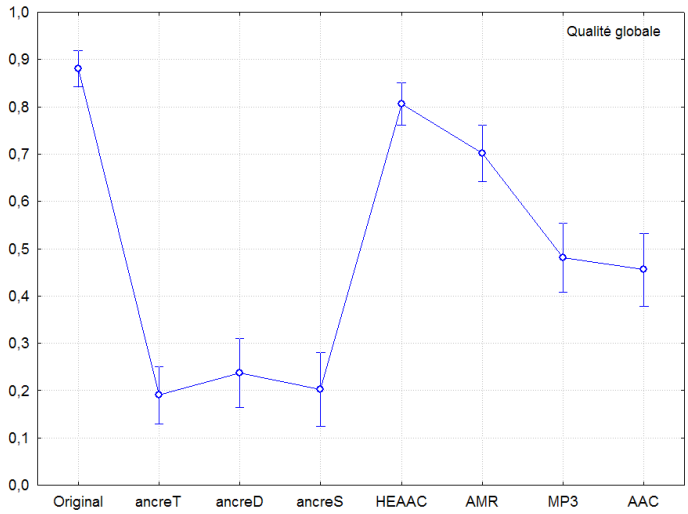


Fig. VI.4 – Moyennes et intervalles de confiance à 95% pour la qualité globale.

Le premier constat est que la version originale a obtenu la moyenne la plus haute avec une valeur de 0.88.

Les trois ancrages "ancreT", "ancreD" et "ancreS" sont notés en basse qualité pour la qualité globale avec des moyennes similaires, respectivement 0.19, 0.24 et 0.20. L'ancrage spatial utilisé dans le chapitre IV qui consistait en une inversion dynamique des canaux droit et gauche, lui, avait été évalué dans la partie haute de l'échelle (figure VI.5).

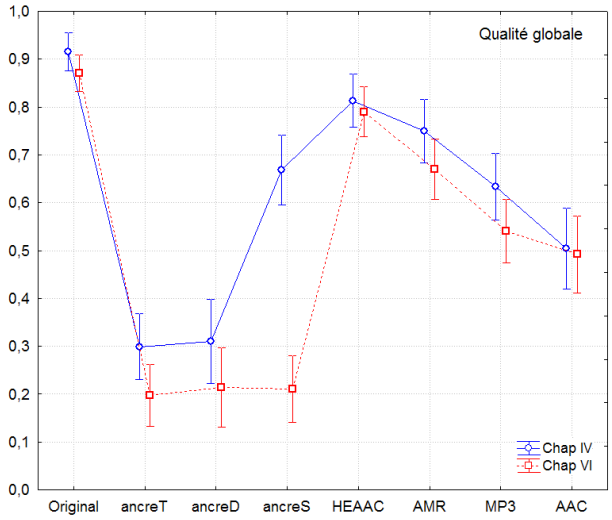


Fig. VI.5 – Moyennes et intervalles de confiance à 95% pour la qualité globale sur les 7 extraits évalués dans les deux tests.

Les quatre codages obtiennent des moyennes similaires à celles obtenues dans le test chapitre IV, en comparant les résultats sur les sept extraits similaires aux deux tests (figure VI.5). Un post-hoc de Tukey permet de préciser que les moyennes de chaque codage obtenues dans le test chapitre IV et le chapitre VI n'ont pas de différences significatives. Concernant la qualité globale, la méthodologie mise en place montre une certaine reproductibilité et fiabilité des résultats. Il s'avère que la modification de l'ancrage spatial n'influence pas la notation de la qualité globale des différentes versions.

VI.3.2 Attributs *Timbre, Espace* et *Défauts*

La deuxième partie du test d'écoute est consacrée à l'évaluation des séquences suivant les trois attributs de qualité *Timbre, Espace* et *Défauts* (figure VI.6).

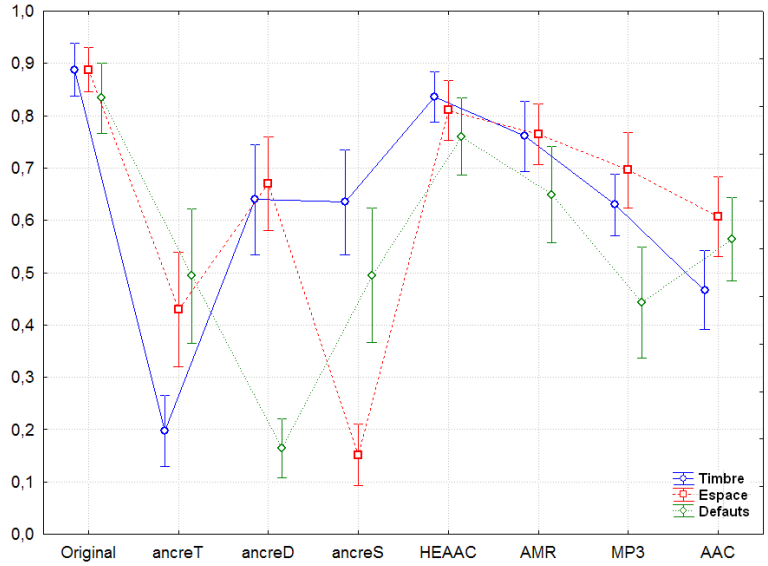


Fig. VI.6 – Moyennes et intervalles de confiance à 95% pour les trois attributs *Timbre, Espace* et *Défauts*.

Pour chacun des trois attributs, la version originale a obtenu le score le plus élevé (>0.83).

Chaque ancrage est l'élément qui obtient la note la plus basse lorsque l'évaluation porte sur l'attribut qui lui est associé, tout en obtenant des notes plus élevées concernant les deux autres attributs. Ainsi, "ancreT" obtient la note de 0.2 pour le *Timbre*, "ancreS" obtient 0.15 pour l'évaluation de l'*Espace* et "ancreD" obtient 0.16 pour l'attribut *Défauts*. Les trois ancrages sont notés en basse qualité <0.2. Chaque ancrage dégrade fortement et principalement l'attribut pour lequel il a été créé.

Concernant les codages "MP3" et "AAC", il est intéressant de noter que, à qualité globale égale, ce sont différents aspects du son qui sont dégradés. Le "MP3' semble plutôt affecter l'attribut *Défauts* tandis que pour le codage "AAC" le *timbre* est l'attribut perceptif le plus dégradé. Ces mêmes observations avaient été relevées dans le chapitre IV et confirment l'intérêt d'une méthode d'évaluation multicritère.

Dans le chapitre IV, le problème soulevé était que la notation de l'ancrage spatial dépendait du contenu. En se focalisant sur l'attribut *Espace*, les notes de l'ancrage spatial inclus dans ce test sont inférieures à 0.28 quel que soit l'extrait évalué (figure VI.7).

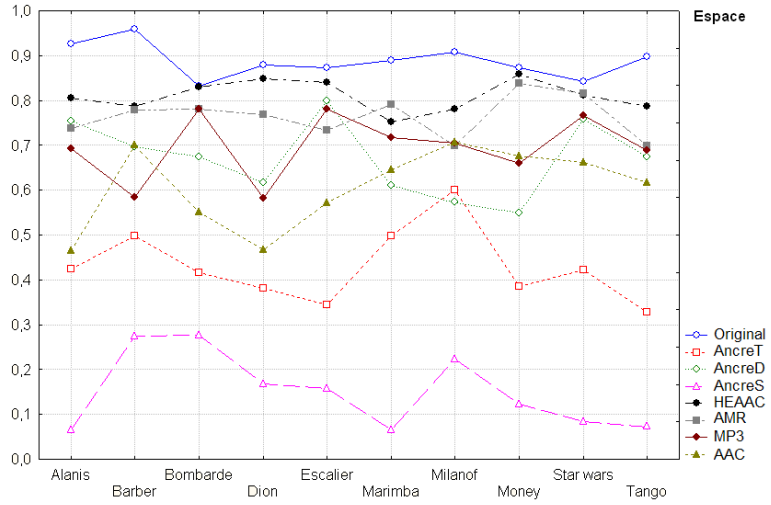


Fig. VI.7 – Moyennes de chaque versions en fonction des extraits pour l'attribut *Espace*.

L'ANOVA montre un effet significatif du test (comparaison chapitres IV et VI) sur la notation de l'attribut *Espace*. Les moyennes des quatre codages sont inférieures deux à deux (figure VI.8), mais sont statistiquement équivalentes par un post-hoc de Fisher.

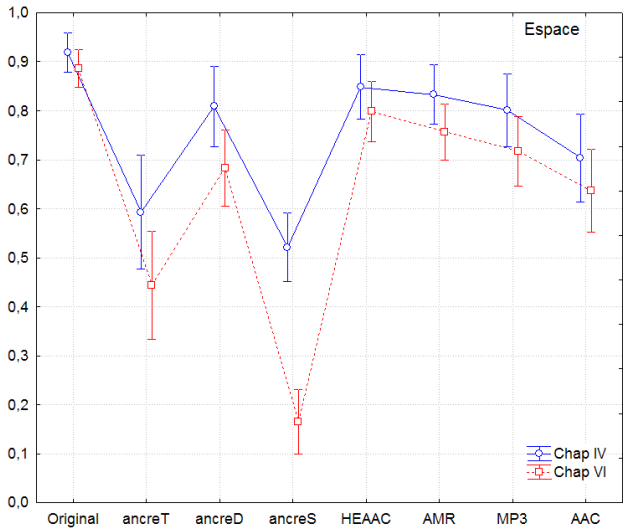


Fig. VI.8 – Moyennes et intervalles de confiance à 95% pour l'attribut *Espace* sur les 7 extraits évalués dans les deux tests.

La modification de l'ancrage spatial interfère principalement sur sa propre notation et impacte peu celles des autres versions notamment les codages. On peut s'interroger sur l'aspect critique de l'ancre spatiale. Ceci peut être expliqué par la nature des codages évalués qui altèrent peu l'aspect spatial des sons. En effet, leurs moyennes sont supérieures à 0.63 concernant l'attribut *Espace*.

VI.3.3 Régression linéaire

Le tableau VI.2 donne les valeurs des corrélations entre les quatre axes d'évaluation de la qualité. Chaque attribut est fortement corrélé à la qualité globale : 0.75 pour le *Timbre*, 0.77 pour l'*Espace* et 0.81 pour les *Défauts*. Avec une valeur de 0.40, les attributs *Espace* et *Défauts* sont les moins corrélés.

Tab. VI.2 – Corrélation entre les quatre attributs de qualité.

Attribut	Timbre	Espace	Défauts
Qualité Globale	0.75	0.77	0.81
Timbre	-	0.58	0.49
Espace	-	-	0.40

D'après les résultats de la régression linéaire, en prenant en compte toutes les versions dont les ancrages, l'attribut *Défauts* possède le poids le plus important sur la qualité globale suivi de l'*Espace* puis du *Timbre*, respectivement 0.52, 0.42 et 0.25 pour les valeurs des coefficients de régression standardisés β. Le poids des attributs *Timbre* et *Espace* est inversé par rapport aux poids obtenus dans le chapitre IV : $\beta = 0.41$ pour le *Timbre* et $\beta = 0.20$ pour l'*Espace*. La différence entre les deux tests réside dans la modification de l'ancrage spatial qui d'après les résultats du paragraphe précédent est plus adapté. Le résultat de la régression linéaire est donc fortement influencé par le choix de l'ancrage spatial.

Les coefficients non standardisés permettent d'établir l'équation de régression :

$$QG = 0.68\ \textit{Défauts} + 0.50\ \textit{Espace} + 0.31\ \textit{Timbre} - 0.39, \tag{VI.1}$$

Le modèle de prédiction permet de prédire 94% de la variance ($R^2 = 0.94$) de la qualité globale en fonction des trois attributs *Timbre*, *Espace* et *Défauts*. L'erreur standard d'estimation est peu importante (0.07). Le modèle semble fiable.

Comme dans le chapitre IV, une régression linéaire est réalisée sur les quatre codages évalués c'est-à-dire sans prendre en compte les ancrages et la version originale. Les valeurs de corrélations entre les attributs sont données dans le tableau VI.3.

Tab. VI.3 – Corrélation en tenant compte uniquement des quatre codages (sans ancres, sans l'original).

Attributs	Timbre	Espace	Défauts
Qualité globale	0.84	0.70	0.86
Timbre	-	0.87	0.65
Espace	-	-	0.55

Les valeurs de corrélation sont équivalentes à celles obtenues dans le test détaillé dans le chapitre IV. Les trois attributs sont corrélés à la qualité globale, et le *Timbre* et l'*Espace* sont fortement corrélés entre eux.

Les valeurs des coefficients de régression standardisés (β) indiquent que les attributs *Timbre* et *Défauts* ont un poids similaire sur la qualité globale respectivement $\beta = 0.551$ et $\beta = 0.546$. Le poids de l'attribut *Espace* est non significatif ($p > 0.5$) avec $\beta = -0.08$.

Comme dans le chapitre IV, les attributs *Timbre* et *Défauts* sont les deux paramètres qui influencent l'évaluation de la qualité globale si on tient compte uniquement des quatre codages.

La principale différence entre les deux tests (chapitres IV et VI) est l'ancrage spatial. Les codages sont identiques. D'après l'analyse de la régression linéaire, la modification de l'ancrage spatial affecte peu l'évaluation des codages.

VI.4 Conclusion

La dégradation appliquée pour l'ancrage spatial consiste à créer un mouvement de balancement entre le canal droit et le canal gauche. Ceci engendre une incohérence dynamique dans la perception de l'espace ainsi que des modifications des caractéristiques spatiales telles que la profondeur, l'enveloppement... Cet ancrage spatial a été testé dans un test d'écoute au casque, basé sur trois attributs de qualité, appliqué à des contenus binauraux. Il obtient la moyenne la plus faible et en basse qualité pour la perception de l'*Espace* et ceci, pour chacun des extraits évalués. De plus, il dégrade faiblement le *Timbre* et les *Défauts*. Cet ancrage répond donc aux exigences et est validé comme ancrage spatial basse qualité à inclure dans un test.

L'attribut *Défauts* reste l'attribut dont l'influence est la plus importante sur la qualité globale. Cependant, de par la modification de l'ancrage spatial, l'*Espace* possède un poids supérieur au *Timbre* sur la qualité globale. La régression linéaire réalisée uniquement sur les quatre codages montre l'influence du *Timbre* et des *Défauts* sur la qualité globale alors que le poids de l'*Espace* est insignifiant.

Les résultats entre ce test et le test réalisé dans le chapitre IV sont similaires et prouvent l'intérêt d'une évaluation multicritère ainsi que la fiabilité de la méthodologie mise en place. Il a été également relevé que le choix de l'ancrage spatial n'impacte pas la notation des codages et donc pose la question de l'intérêt réel d'un tel ancrage dans un test d'écoute.

Chapitre VII

Un unique ancrage triplement dégradé

Trois ancrages liés à trois dimensions des sons ont été inclus dans une méthode de test audio dédiée à l'évaluation des sons spatialisés. Ainsi, après l'élicitation de ces trois familles d'attributs qui sont le *Timbre*, l'*Espace* et les *Défauts*, trois ancrages spécifiques à chacun de ces attributs ont été validés et inclus dans des tests d'écoute. Cependant, à terme, le but est d'évaluer des systèmes de codages ou de restitution et le nombre de stimuli à évaluer étant limité, l'intégration de trois ancrages dans un test diminue leur nombre. En effet, l'évaluation de plus de dix versions simultanément s'avère difficile et laborieuse. Dans ce chapitre, le test inclut un ancrage unique qui est la superposition des ancrages spécifiques et qui dégrade donc les trois axes perceptifs simultanément.

VII.1 L'ancrage unique

L'ancrage unique est la superposition des trois ancrages spécifiques à chaque attribut proposés lors des tests précédents (chapitre IV et VI). Il s'agit d'ajouter un bruit rose et une série de clics sur la version originale (chapitre IV paragraphe 1.2) puis d'appliquer un mouvement oscillant entre le canal droit et le canal gauche (chapitre VI paragraphe 1) et enfin de réaliser un filtrage passe bas butterworth d'ordre 8 avec une fréquence de coupure à 3.5 kHz. La dégradation du *Timbre* et de l'*Espace* engendre une perte d'information sur le signal, la dégradation *Défauts* est donc appliquée en premier. La dégradation de l'*Espace* est réalisée en second. Le filtrage passe-bas est appliqué en dernier, après tous les traitements, pour la dégradation du *Timbre*.

VII.2 Protocole expérimental

Dix-huit testeurs participent à ce test d'écoute. Les conditions sont identiques aux tests réalisés dans les chapitres précédents c'est-à-dire une écoute au casque dans une cabine insonorisée (chapitre IV).

VII.2.1 Stimuli

Les extraits présentés lors du test sont les dix extraits utilisés dans les chapitre V et VI (tableau V.1).

Six versions sont proposées pour chacun des extraits (tableau VII.1). Les quatre codages sont les mêmes que ceux utilisés dans le chapitre VI.

Tab. VII.1 – Description des versions évaluées.

Original	La référence : le fichier original
HEAAC	HE-AACv2 à 40 kbits/s
AMR	AMR WB+ à 48 kbits/s
MP3	MP3 à 64 kbits/s
AAC	AAC à 32 kbits/s
anchor	Ancrage triplement dégradé

VII.2.2 Déroulement du test

Le test suit la même procédure que celle des chapitres précédents. Toutes les versions d'un extrait sont présentées sur une même interface. Les auditeurs écoutent les séquences et notent la qualité perçue sur une échelle comportant les labels "Basse qualité" et "Haute qualité" respectivement pour chaque extrémité. Au moins une version doit être obligatoirement notée au maximum de l'échelle. Lors de la première phase de test, il est demandé aux auditeurs d'évaluer la qualité globale des séquences. Dans la seconde phase de test, ils doivent évaluer chacun des trois attributs proposés *Timbre*, *Espace* et *Défauts*. Les consignes de test sont décrites dans l'annexe A.3 et A.4. Les trois attributs sont présentés simultanément sur une seule interface (figure VI.3, chapitre VI).

VII.3 Résultats

VII.3.1 Qualité globale

La figure VII.1 représente les moyennes et intervalles de confiance à 95% obtenus pour l'évaluation de la qualité globale.

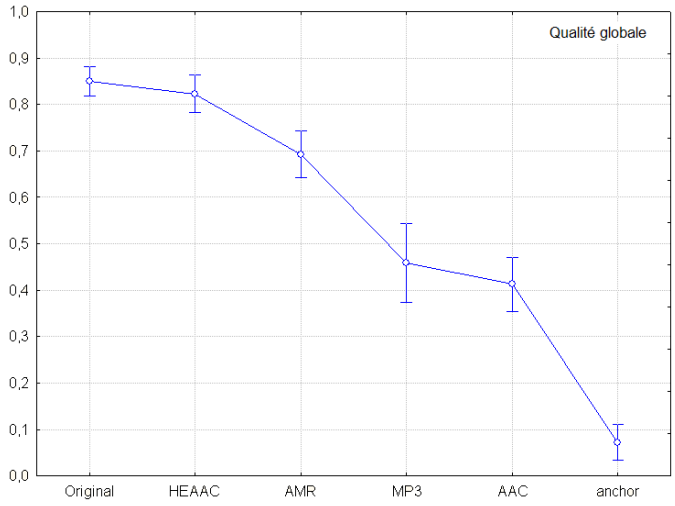


Fig. VII.1 – Moyennes et intervalles de confiance à 95% pour la qualité globale.

L'original obtient la meilleure moyenne (0.85). Cependant, avec une moyenne de 0.82, le codage "HEAAC" lui est statistiquement équivalent selon un post-hoc de Tukey. La consigne de mettre une version au maximum de l'échelle étant respectée, l'absence de référence explicite permet d'évaluer un objet à une qualité équivalente voir supérieure à la version originale non dégradée et d'évaluer la qualité et non la fidélité par rapport à une référence. Selon Tukey, l'"AMR" se différencie de toutes les autres versions avec une moyenne de 0.69 et les versions "MP3" et "AAC" sont équivalentes avec respectivement 0.45 et 0.41 de moyenne. L'ancrage unique obtient une moyenne très basse de 0.07. Ces résultats restent cohérents avec les résultats obtenus lors du test avec les trois ancres spécifiques (chapitre VI). Les moyennes obtenues par chaque codec sont statistiquement équivalentes selon le test de Tukey (figure VII.2).

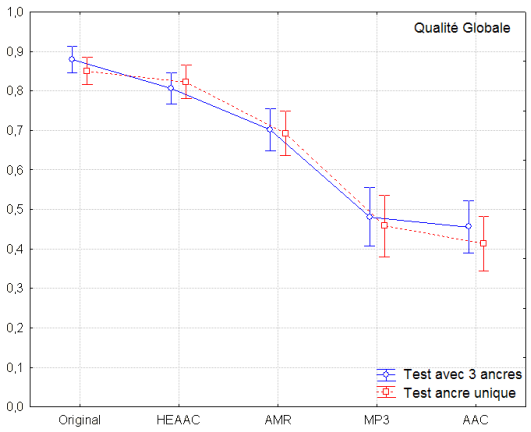


Fig. VII.2 – Moyennes et intervalles de confiance à 95% pour la qualité globale en comparaison les résultats du chapitre VI.

VII.3.2 Attributs *Timbre*, *Espace* et *Défauts*

Les résultats concernant l'évaluation simultanée des trois attributs de qualité sont représentés dans la figure VII.3.

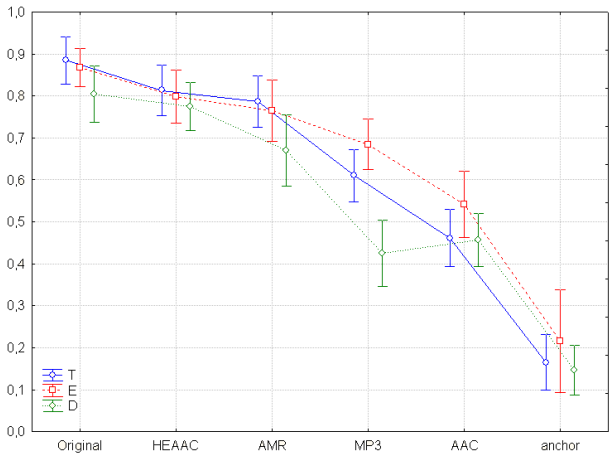


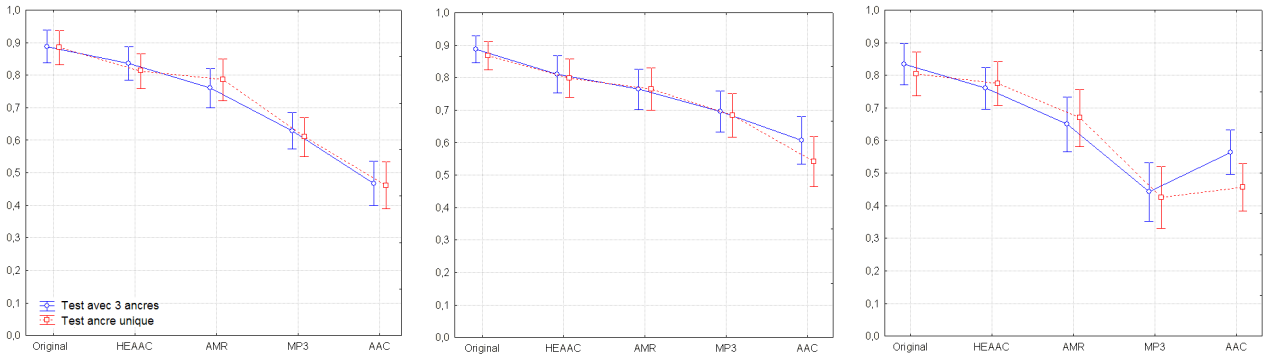
Fig. VII.3 – Moyennes et intervalles de confiance à 95% pour les trois attributs.

Pour chaque attribut, l'ordre de qualité des versions est identique : "Original", "HEAAC", "AMR", "MP3", "ACC", "anchor". Excepté pour l'attribut *Défauts*, l'"AAC" obtient une moyenne légèrement supérieure à celle du "MP3", respectivement 0.46 et 0.43.

L'original obtient la meilleure moyenne pour chacun des attributs et se différencie des cinq autres versions considérant un post-hoc LSD de Fisher pour le *Timbre* et l'*Espace*, tandis que pour l'attribut *Défauts*, l'"Original" est statistiquement similaire à "HEAAC". Pour chaque attribut, les différentes versions sont bien discriminées entre elles, seules les versions "HEAAC" et "AMR" pour le *Timbre* et l'*Espace* et le "MP3" et l'"AAC" pour l'attribut *Défauts* sont statistiquement équivalentes au sens de Fisher. L'ancrage "anchor" est noté en basse qualité pour les trois attributs (*Timbre* :0.16 ; *Espace* :0.22 ; *Défauts* :0.15). Il dégrade de manière équivalente la qualité des trois axes de perception proposés.

A partir du post-hoc de Fisher, il y a pas de différences significatives en fonction des attributs pour les versions "Original" et "HEAAC". En revanche, pour la version "AMR", la moyenne de *Défauts* se différencie des deux autres attributs eux-mêmes statistiquement similaires. Les moyennes de "MP3" sont, chacune, significativement différentes. Pour l'"AAC", les moyennes du *Timbre* et des *Défauts* sont équivalentes et se différencient, toutes deux, de l'*Espace*. Ceci montre que le test multicritère a du sens.

Les résultats obtenus sont similaires à ceux analysés avec la présence de trois ancres dans le chapitre VI (figure VII.4) à l'exception de la version "AAC" laquelle a obtenu une moyenne inférieure pour l'*Espace* et les *Défauts* avec l'ancre unique.



(a) Timbre. (b) Espace. (c) Défauts.

Fig. VII.4 – Comparaison des moyennes et intervalles de confiance à 95% avec les résultats du chapitre VI.

L'utilisation de trois ancres spécifiques ou d'une ancre triple donne les mêmes résultats sur l'évaluation des codages.

VII.3.3 Régression linéaire

Le tableau VII.2 donne les valeurs des corrélations entre les quatre axes d'évaluation de la qualité. Tous les attributs sont fortement corrélés entre eux.

Tab. VII.2 – Corrélation.

Attribut	Timbre	Espace	Défauts
Qualité Globale	0.96	0.93	0.97
Timbre	-	0.97	0.92
Espace	-	-	0.89

Cette colinéarité des attributs rend l'étude de la régression linéaire difficilement analysable. Les variables mesurent, en fait, la même chose. Le modèle suggéré donne le poids le plus important à *Défauts* sur la qualité globale suivi du *Timbre* puis de *Espace* respectivement 0.56, 0.43 et 0.02 pour les valeurs des coefficients de régression standardisés β. L'attribut *Espace* serait un facteur négligeable ce qui coïncide avec les résultats de la régression linéaire réalisée sur les codages uniquement dans les chapitres IV et VI. Les régressions linéaires des tests avec trois ancres (chapitres IV et VI) et avec ancre unique (chapitre VII) ne peuvent être comparés tant le nombre élevé d'ancres modifient les résultats. Par conséquent, une régression linéaire est réalisée sur les codages uniquement, sans prendre en compte les ancrages et l'original.

Le tableau VII.3 donne les valeurs de corrélations entre les attributs calculées sur les quatre codages uniquement.

Tab. VII.3 – Corrélation.

Attribut	Timbre	Espace	Défauts
Qualité Globale	0.88	0.79	0.92
Timbre	-	0.90	0.81
Espace	-	-	0.73

Tous les attributs sont corrélés entre eux. Comme expliqué dans le paragraphe précédent, la colinéarité rend les résultats de la régression linéaire difficilement exploitable. La comparaison des coefficients β peut s'effectuer seulement lorsque la régression linéaire est faite sans prendre en compte les trois ancrages et l'original. Le tableau VII.4 regroupe les coefficients β obtenus dans les trois tests pour une analyse sur les quatre codages uniquement.

Tab. VII.4 – Comparaison des coefficients β obtenus chapitre IV, VI et VII en prenant en compte uniquement les codages.

Coefficient β	Timbre	Espace	Défauts
Chapitre IV (3 ancrages)	0.59	0.03	0.44
Chapitre VI (3 ancrages avec ancre spatiale adaptée)	0.551	-0.08	0.546
Chapitre VII (ancrage unique)	0.44	-0.05	0.60

Pour l'évaluation des quatre codages, le *Timbre* et les *Défauts* sont des facteurs qui expliquent la variabilité de la qualité globale, l'*Espace*, lui, est négligeable. Toutefois, les résultats de la régression sont à prendre avec précaution étant donné la colinéarité des attributs.

VII.4 Conclusion

L'ancrage étudié consiste à sommer trois dégradations spécifiques à chaque attribut de qualité évalué *Timbre, Espace* et *Défauts*. Ces trois altérations sont l'ajout de bruit et de clics, l'application d'un mouvement oscillant (gauche, droit) et d'un filtrage passe-bas. Le remplacement des trois ancrages spécifiques par l'ancre unique affecte peu la notation de la qualité perçue par les auditeurs. Les résultats des tests précédents et de ce test sont similaires en terme de qualité globale et pour les trois attributs. En revanche, l'ancre unique a l'avantage de réduire le test de deux ancrages et donc permet d'inclure des versions supplémentaires à évaluer, par exemple des codecs. De plus, le test montre que cette ancre dégrade les trois axes de qualité de manière équivalente et en basse qualité.

Les résultats sont similaires cependant tous les attributs sont corrélés entre eux. Si on compare les régressions linéaires sans prendre en compte les ancrages, l'*Espace* semble avoir peu d'influence sur la qualité globale.

Chapitre VIII

La méthodologie

La recommandation ITU-R BS.1534 (2003) (MUSHRA) est la méthode utilisée pour évaluer les sons de qualité intermédiaire. Cependant, Zielinski *et al.* (2007b) ont révélé que cette méthode comporte certains biais et notamment pour les sons spatialisés. Les études détaillées dans les chapitres précédents chapitre (II-VII) ont permis d'élaborer une méthodologie pour l'évaluation de la qualité des sons spatialisés. Tout en gardant les bénéfices de MUSHRA tels qu'une comparaison multiple, une référence cachée, des ancrages, le choix des sujets et la longueur des stimuli, la méthode propose d'autres attributs de qualité, une échelle de qualité différente et des ancrages adaptés.

VIII.1 Les attributs de qualité

La méthode propose une évaluation de la qualité basée sur une analyse multicritère (chapitre II). Quatre attributs sont intégrés comme axes perceptifs pour juger la qualité sonore :

- Qualité globale : qualité globale du son, tous paramètres confondus.
- Timbre : couleur du son, e.g. brillance, coloration, clarté, dureté, couleur du timbre, équalisation, richesse...
- Espace : fait référence à l'impression relative aux caractéristiques spatiales, e.g. profondeur, largeur, localisation, distribution spatiale, réverbération, spatialisation, distance, enveloppement, immersion...
- Défauts : artefacts ou nuisances présents dans le son, e.g. bruit, distorsion, bruit de fond, sifflement, bourdonnement, coupure...

VIII.2 Caractéristiques générales de la méthode

VIII.2.1 Sujets

Les auditeurs sont considérés comme "experts" comme le préconise la recommandation ITU-R BS.1534 (2003). Ils ont acquis de l'expérience en écoute et en détection de dégradations dans des stimuli audio. Une vingtaine d'auditeurs est requise pour un test audio.

VIII.2.2 Stimuli

La longueur des séquences évaluées doit être comprise entre 10 et 20 secondes (ITU-R BS.1534, 2003). Les extraits choisis doivent couvrir une palette diversifiée de contenus

telle que de la musique, des extraits de films, de la parole, des sons d'environnements... La méthode étant dédiée aux sons spatialisés, les extraits doivent présenter une spatialisation prononcée.

Pour chaque extrait, le test inclut deux versions, le signal original et un ancrage, en plus des codages évalués.

Des dégradations spécifiques à chaque attribut ont été définies.

La dégradation du *Timbre* est un filtrage de la version originale avec un filtre passe-bas butterworth d'ordre 8 avec une fréquence de coupure à 3.5 kHz.

La conception de *Défauts* consiste à ajouter un bruit rose (RBS≈ 30dB) ainsi qu'une série de clics sur la version originale.

La détérioration de l'*Espace* consiste à appliquer un mouvement de rotation du signal mono (somme de tous les canaux) d'une période de 6.54 secondes à l'aide d'une fonction trapèze avec un temps d'arrêt à droite et à gauche de 1.93 secondes.

L'ancrage utilisé est un ancrage unique qui est la somme de ces trois dégradations à appliquer dans l'ordre suivant : l'ajout de bruit et clics, le mouvement de rotation puis le filtrage.

VIII.2.3 L'échelle de notation

L'échelle employée est une échelle de qualité avec pour extrémités les étiquettes "Basse qualité" et "Haute qualité" sans label intermédiaire (Figure VIII.1) (Le Bagousse *et al.*, 2012; Marins *et al.*, 2008). Les auditeurs n'ont pas de repères numériques. En revanche pour l'analyse des résultats, l'échelle est convertie en échelle numérique de 0 ("Basse qualité") à 1 ("Haute qualité") par pas de 0.01 (ITU-R BS.1534, 2003).

Fig. VIII.1 – L'échelle de qualité

VIII.2.4 Le protocole de test

Le test se déroule en deux parties. La première concerne l'évaluation de la *qualité globale* et la seconde celle des attributs de qualité *Timbre*, *Espace* et *Défauts* évalués simultanément.

Basée sur la recommandation ITU-R BS.1534 (2003), les évaluations reposent sur des tests en comparaison multiples. Toutes les versions d'un même extrait sont présentées simultanément (A, B, C ...) comme le montre la figure VIII.2.

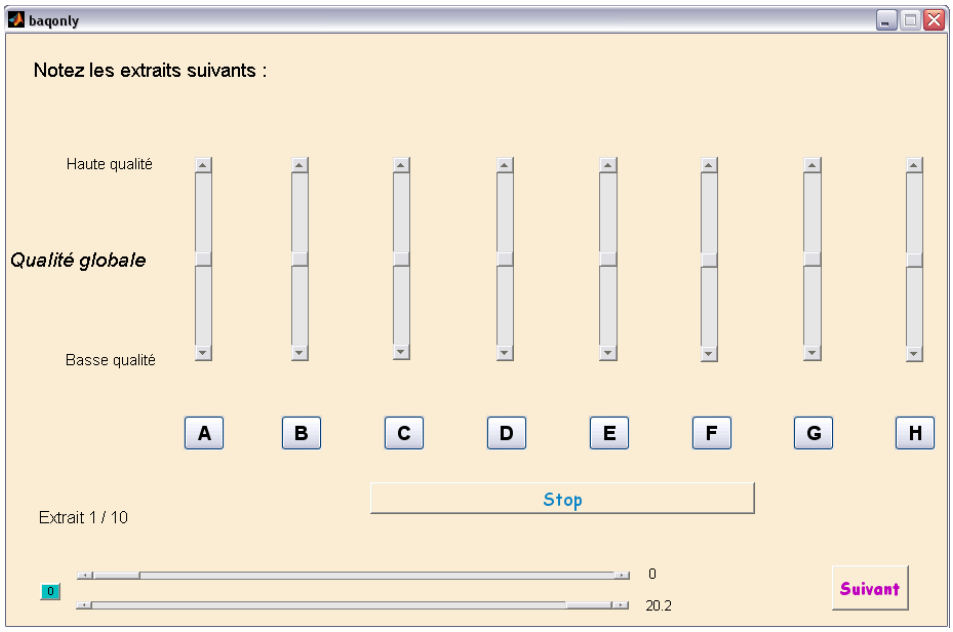


Fig. VIII.2 – Interface de test pour l'évaluation de la *qualité globale*

Le test ne présente pas de référence explicite : la version originale d'un extrait est incluse comme référence cachée. L'absence de référence explicite permet de supprimer la tache de reconnaissance de celle-ci et de remplacer l'évaluation de fidélité par une évaluation de qualité. Elle évite également d'imposer une référence haute qualité aux auditeurs. De cette manière, en particulier dans le contexte des sons spatialisés, les objets évalués ne sont pas comparés directement avec cette référence et peuvent obtenir une note plus élevée que celle-ci. Par exemple, l'utilisation d'un stéréo enhancer peut donner aux auditeurs l'impression d'une amélioration de la qualité de l'espace.

Les auditeurs écoutent les différentes versions autant de fois que nécessaires. Ils ont la possibilité de réduire la durée d'écoute en utilisant les curseurs horizontaux, après avoir écouté l'extrait au moins une fois en intégralité. Ensuite, ils notent chaque version évaluée sur l'échelle de qualité. La consigne donnée aux auditeurs est que pour chaque extrait, la version perçue de meilleure qualité doit être notée au maximum de l'échelle. Une fois leur évaluation terminée, ils passent à l'extrait suivant avec le bouton "suivant".

Concernant la seconde partie du test, l'évaluation des trois attributs se fait simultanément sur une même interface (figure VIII.3).

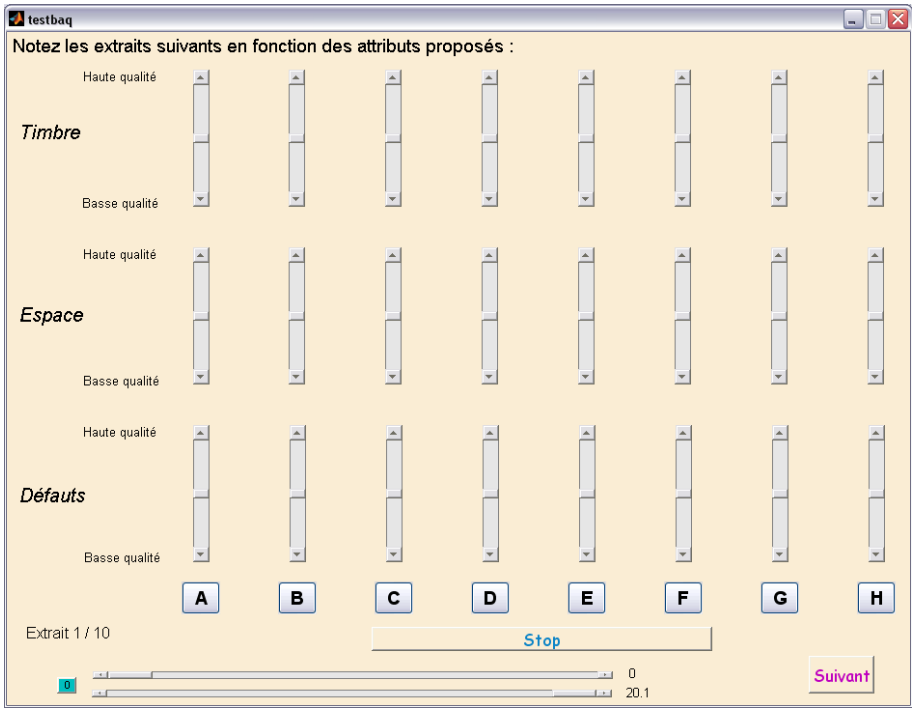


Fig. VIII.3 – Interface de test pour l'évaluation du *Timbre*, de l'*Espace* et des *Défauts*

VIII.3 Conclusion

Cette méthode de tests subjectifs est basée sur l'évaluation de trois attributs représentant chacun un axe perceptif de la qualité. Elle s'applique à l'évaluation des sons spatialisés de qualité intermédiaire. Un ancrage unique, basse qualité, a été inclus incorporant des dégradations sur les trois attributs perceptifs. La méthode ne propose pas de référence explicite. Elle permet de mettre en évidence le ou les axes de qualité qui sont altérés par les systèmes testés, par exemple des codages.

Conclusion

Une méthodologie de tests audio a été développée pour l'évaluation subjective de la qualité des sons spatialisés pour pallier les limites des méthodes actuelles, notamment celles de l'union internationale des télécommunications (ITU), l'ITU-R BS.1116 et l'ITU-R BS.1534.

L'évaluation basée sur la qualité globale est indispensable, mais reste insuffisante pour exprimer l'impact des dégradations sur la perception d'un signal audio. Le but des travaux présentés ici était, dans un premier temps, de définir des axes de perception de la qualité en prenant en compte, désormais, la dimension spatiale du son dans le contexte de l'évaluation de systèmes de codages. Un large nombre d'attributs a été élicité dans la littérature scientifique et il est impossible d'inclure tous ces termes dans un test d'écoute. Des groupements d'attributs ont donc été formés pour établir des catégories perceptives de la qualité. Trois attributs généraux ont été formés : le *Timbre*, l'*Espace* et les *Défauts*.

Les attributs étant établis, ils ont été inclus dans un test d'écoute réalisé sur un système 5.1. Il a été demandé à des auditeurs "experts" d'évaluer dans un premier temps la qualité globale puis les trois attributs. Deux méthodes de présentation de ces trois attributs ont été expérimentées : l'évaluation de chacun des attributs successivement et l'évaluation simultanée des trois attributs. Les modes ne présentaient pas de différence significative en terme de résultat. Cependant, la présentation simultanée a l'avantage de diminuer le temps de test et donc la fatigue de l'auditeur. Un ancrage basse qualité spécifique à chaque attribut a été inclus dans le set de stimuli. Pour le *Timbre* et les *Défauts*, l'ancrage a bien été identifié, en revanche, l'ancrage spatial n'a pas été reconnu en basse qualité. L'échelle utilisée était une échelle de notation sans grade intermédiaire avec pour extrémités les termes "basse qualité" et "haute qualité". Le test a été réalisé sans référence explicite mais la version originale était incluse comme référence cachée. En effet, l'intérêt était de privilégier une évaluation de qualité inter-objets plutôt qu'une évaluation de fidélité par reconnaissance de cette référence et comparaison des objets avec celle-ci. De plus, l'absence de référence explicite laissait libre la notation en haute qualité aux auditeurs. Un objet pouvait alors obtenir une note supérieure à celle de la dit "référence". Les auditeurs avaient la consigne de noter au maximum de l'échelle au moins une version dans le but de garder une dynamique constante de l'échelle de notation quel que soit l'extrait évalué et pour une reproductibilité du test d'écoute.

La méthode a, ensuite, été appliquée à l'évaluation de contenus binauraux restitués au casque afin de vérifier la pertinence de la méthode sur un autre système de spatialisation. L'évaluation multicritère a montré son intérêt pour des systèmes de codages de moyennes et fortes dégradations qualifiées de qualité intermédiaire par la méthode ITU-R BS.1534. En effet, des différences significatives ont été observées entre les moyennes de chaque at-

tribut pour un système de codage donné. L'ancrage spatial consistait en une inversion du canal droit et du canal gauche. Il a été évalué en moyenne au milieu de l'échelle de qualité. Cependant, suivant le contenu de la séquence, il a été évalué en basse ou en haute qualité.

Un travail spécifique a été réalisé sur la conception d'un ancrage spatial adapté à tout contenu et évalué en basse qualité. Le procédé validé a consisté à appliquer un mouvement oscillant de gauche à droite du signal mono qui crée une incohérence dynamique de l'image spatiale.

Pour chaque ancrage, la dégradation de chacun des attributs et uniquement cet attribut a été validée lors d'un test d'écoute. Cependant, le fait d'inclure trois ancrages limitait le nombre de versions à évaluer. Un ancrage unique, étant le mélange des trois dégradations spécifiques, s'est révélé aussi efficace que trois ancres dédiées. Cependant, l'évaluation des codages est indépendante du choix des ancrages. Les codages ont obtenu des notes équivalentes quelques soit l'ancre spatiale utilisée ou avec une seule ancre triplement dégradée. La présence d'ancrages dans un test d'écoute est donc mise en cause notamment sur sa pertinence pour l'évaluation des objets du test. Un test sans ancre pourrait alors être mis en place. Notons tout de même que leur rôle premier est de permettre la reproductibilité des tests et l'utilisation d'une large dynamique de l'échelle de notation.

Les différents tests ont montré que l'attribut *Défauts* était l'élément qui influençait majoritairement la qualité globale. La régression linéaire, en prenant en compte l'analyse des résultats sur les codages uniquement (sans les ancres et sans l'original), a montré que la variabilité de la qualité globale était expliquée par le *Timbre* et les *Défauts* et que l'*Espace* était une variable négligeable. Cependant, la corrélation importante entre les attributs remet en question la validité de la régression linéaire.

La méthode d'évaluation subjective de la qualité pour les sons spatialisés a été utilisée pour deux systèmes de restitution spatialisée, le 5.1 et le binaural. Dans le but d'avoir une méthode recommandée valide pour tout système spatialisé, la méthode pourra être utilisée avec d'autres techniques de spatialisation tel que Ambisonics ou Wave Field Synthesis. L'ancrage unique notamment pour la dégradation de l'*Espace* pourra alors être adapté.

Annexe A

Les consignes données aux auditeurs

Les consignes de test sont données à chaque auditeur au début de chaque session.

A.1 Consignes de test pour la MDS

Votre tâche est de noter sur l'échelle indiquée la similarité entre deux attributs de la liste. Pour le faire vous disposez d'une interface. Vous avez la possibilité de placer le curseur sur la totalité de l'échelle, celle-ci allant de "très différents" à "très similaires". Toutes les paires vont vous être présentées successivement. Cliquez sur suivant, lorsque la notation est terminée (vous ne pourrez pas revenir en arrière). Dans le cas où la définition d'un des termes proposés vous semble incertaine, vous devez vous fier à votre propre interprétation.

Quelques explications sur la notion de similarité dans le cadre de ce test. Vous devez juger la similarité comme l'appartenance à une même caractéristique sonore. Prenons des exemples : L'orange et le citron n'ont pas le même goût ni la même forme mais ils appartiennent à la famille des agrumes, de même que le bleu et le jaune sont des couleurs.

A.2 Consignes de test pour la catégorisation libre

Une liste de 28 attributs utilisés pour qualifier le son est proposée dans la première colonne de l'interface. Pour ce test, votre tâche est de regrouper ces attributs en catégories. Il vous suffit de cliquer sur l'attribut et de le déplacer avec la souris. Vous êtes libre dans la formation et le choix du nombre de ces catégories avec pour instruction un nombre minimum de 2 et un maximum de 5 catégories. Chaque groupe ainsi formé peut contenir un nombre différent d'attributs. Dans le cas où la définition d'un des termes de la liste vous semble incertaine, vous ne devez pas chercher sa définition, et devez vous fier à votre propre interprétation. Cependant, veuillez l'indiquer dans la partie commentaire. Une fois que vous avez composé les différentes familles, vous devez leurs attribuer un nom qui les caractérise (dans les cadres oranges de l'interface).

A.3 Consignes de test pour l'évaluation de la qualité globale

Le but du test est de juger la qualité audio globale d'extraits sonores.

Vous allez entendre x extraits différents. A chaque extrait correspond une interface de test vous permettant d'écouter y versions différentes (A, B, C, D...) de cet extrait.

Vous pouvez écouter les versions autant de fois que vous le souhaitez, et/ou sélectionner des passages de l'extrait à réécouter. Il vous est cependant demandé d'écouter toutes les versions au moins une fois en entier.

En vous appuyant sur leur comparaison, vous devez placer les y versions sur l'échelle continue définie par les extrémités "haute qualité" et "basse qualité". Pour chaque extrait, vous devez positionner la version sonore que vous considérez comme étant de plus haute qualité au maximum de l'échelle.

Vous devez juger la qualité des extraits en fonction du critère de qualité globale du son.

A.4 Consignes de test pour l'évaluation des attributs

Le but du test est de juger la qualité d'extraits sonores selon les critères de timbre, d'espace et de défauts.

Vous allez entendre x extraits différents. A chaque extrait correspond une interface de test vous permettant d'écouter y versions différentes (A, B, C, D...) de cet extrait.

Vous pouvez écouter les versions autant de fois que vous le souhaitez, et/ou sélectionner des passages de l'extrait à réécouter.

En vous appuyant sur leur comparaison, vous devez placer les y versions sur l'échelle continue définie par les extrémités "haute qualité" et "basse qualité", et ce pour chaque critère : timbre, espace et défauts.

* Timbre : caractérise la couleur du son (brillance, coloration, clarté, dureté, couleur du timbre, équalisation, richesse...).

* Espace : fait référence à l'impression spatiale relative aux caractéristiques spatiales (enveloppement, localisation, réverbération, distance, immersion, largeur, spatialisation, distribution spatiale du son, profondeur...).

* Défauts : sont les artefacts ou nuisances présents dans le son (bruit, distorsion, bruit de fond, sifflement, bourdonnement, coupure...).

Pour chaque extrait et chaque critère, vous devez positionner la version sonore que vous considérez comme étant de plus haute qualité (selon ce critère) au maximum de l'échelle.

Annexe B

Liste des articles - congrès

- S. Le Bagousse, C. Colomes et M. Paquier, State of the Art on Subjective Assessment of Spatial Sound Quality, AES Int. Conf. on Sound Quality Evaluation, Suède, Pitea, 2010.

- S. Le Bagousse, M. Paquier et C. Colomes, Families of Sound Attributes for Assessment of Spatial Audio, AES 129th Convention, USA, San Francisco, 2010.

- S. Le Bagousse, M. Paquier, C. Colomes et Samuel Moulin, Sound Quality Evaluation based on Attributes - Application to Binaural Contents, AES 131th Convention, USA, New York, 2011.

- S. Le Bagousse, M. Paquier et C. Colomes, Assessment of spatial audio quality based on sound attributes, Acoustic's 2012, France, Nantes, 2012.

STATE OF THE ART ON SUBJECTIVE ASSESSMENT OF SPATIAL SOUND QUALITY

SARAH LE BAGOUSSE[1], CATHERINE COLOMES[1], AND MATHIEU PAQUIER[2]

[1] *Orange Labs France Télécom R&D, Cesson Sévigné, France*
sarah.lebagousse@orange-ftgroup.com
catherine.colomes@orange-ftgroup.com
[2] *Laboratoire d'Informatique des Systèmes Complexes (LISyC), Brest, France*
Mathieu.Paquier@univ-brest.fr

A new aim of sound technologies is spatial reproduction, which raises new questions about their quality assessment. This literature review deals with spatial audio quality: for audio coding, assessment is made through use of two mainly subjective ITU-R test methods. But, they are restricted to the evaluation of the overall quality. The finding, through various studies, of some features specific to surround sound drove us to wonder whether they can be included in a new quality assessment.

INTRODUCTION

For the new technologies around spatial sound, the challenge, nowadays, is to lead to realistic sensations. A prerequisite to the broadcasting of such audio contents is a careful evaluation of their quality. Two "well known" subjective test methods recommended by the International Telecommunication Union (ITU) are currently used to evaluate the perceived quality of audio coding. But, none of them takes into account the spatial dimension of sound. Indeed, the result is a global judgment on a single axis called 'basic audio quality' (BAQ). Concerning surround sound, different criteria specific to it have been highlighted over the last years. These considerations drove us to carry out an overall study aimed at developing a new methodology for subjective assessment of surround sound quality. A prerequisite was to gain more insight into the perceptual features of spatial sound and into their degree of involvement in the assessment of audio quality so as to further include some of them in a new test methodology. This literature review is a preliminary to this overall study. It presents, at first, the different surround sound attributes and then some test methods in use in assessment of audio quality.

1 PERCEPTUAL ATTRIBUTES OF SOUND

According to Bech [1], an auditory attribute is "a perceptual characteristic of a sound stimulus". Numerous investigations have been devoted to the finding and listing of audio quality perceptual features. Several elicitation methods have been employed to identify all of these criteria. The available attributes are classified into two families: timbral and spatial. The latter has become of key importance with the increase of surround sound systems.

1.1 Elicitation methods

The technique employed by Berg and Rumsey to look for the sound spatial attributes is the *repertory grid technique* (RGT) [2]. It comes from the earlier study by Kelly [3] and relies on personal constructs, the perception and the language of everyone in order to avoid the limits generated by researchers influence or by imposed scales. This procedure can be split into three main steps [4]:

. Elicitation phase: subjects are asked to compare three stimuli and to indicate the two of them they feel as the most similar, and thus, different from the third. Then, by using two antonym terms, they describe the perceived similarity and difference.

. Scaling phase: The previous two labels are considered as the endpoints of a bipolar construct scale. Then stimuli are scored along it. A matrix grid per person is produced with the scales and the score for each excerpt (Fig. 1)

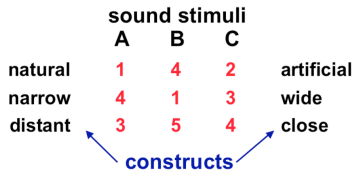


Figure 1: A grid for a subject with scale examples [4]

. Data analysis: through use of a clusters analysis or a principal component analysis (PCA).

OPAQUE (Optimisation of Perceived Audio quality Evaluation) is a software that automatically runs the RGT procedure to ease the elicitation and grading of

audio qualities [5]. A weakness of RTG is that a difference between two sounds can be missed when both are always assimilated to a third sound more dissimilar [6].

Besides RTG, Choisel and Wickelmaier used the method *perceptual structure analysis* (PSA) to assess multichannel reproduced sounds [6]. The method strength is its mathematical foundation based on formal concept analysis [7] and on knowledge space theory [8]. Briefly, the total set of sounds corresponds to a domain denoted by X and σ is a collection of subsets of X, corresponding to auditory attributes; $\langle X , \sigma \rangle$ forms a perceptual structure [6] [9]. A lattice graph represents the domain and their subsets, the nodes denote characteristics shared by sounds. For the test progress, assessors are asked, at first, to listen to three excerpts and to answer by 'Yes' or 'No' to the question "Do sounds 'a' and 'b' share a feature which 'c' does not have?". The highest interest is that this first identification of features is made without asking subjects to put a name on it. After analysis of perceptual structures, the subjects label the common features and write a short description of them.

The *MultiDimensional Scaling* (MDS) is an indirect elicitation method [10]. The aim is to reveal various dimensions and 'hidden meaning' in the data [11]. It is based on the dissimilarity between a pair of items. Assessors are asked to compare two stimuli and to score them on a dissimilarity scale. For example, 0 is 'very similar' and 1 is 'very dissimilar'. The analysis is the interpretation of dissimilarities as distances between the presented sounds represented in a multidimensional graph. Then the extracted dimensions must be interpreted. To evaluate a set of car sounds, Susini and McAdams made a multidimensional scaling (MDS) with CLASCAL program [12].

Guastavino and Katz used *verbal transcript* to highlight sound features. In this experiment, the subjects had to describe the soundscapes with a free verbalization task [13]. The analysis of the data was a semantic analysis on the spontaneous descriptions. Synonyms and linguistic devices contructed on the same stem were grouped into semantic themes that allowed one to identify attributes.

The asset of all of these methods is the lack of predefined dimension, and to be more precise, the lack of imposed quality attributes. Assessors are free in their judgment.

1.2 Timbral attributes

According to the American Standards Association, timbre is defines as "[...] that attribute of sensation in terms of which a listener can judge that two sounds having the same loudness and pitch are dissimilar". Timbre is defined through two different approaches [14]: the causal timbre corresponds to the identification, for example, identifying a singer or recognizing a music instrument. This is the first reaction by a listener at the listening of a sound. The second approach is the qualitative aspect. It is the second reaction after the identification. But it is correlated to individual knowledge. It is difficult to judge or even to identify something unknown. Among the timbral features highlighted by earlier studies on audio quality and timbral attributes [15], let us cite, for example, brightness, clearness, softness [16], fullness, fidelity, pleasantness [17], naturalness [18], richness, tone color, hardness, emphasis [19], timbre, coloration [13]. (See Table 1)

1.3 Spatial attributes

It is essential to distinguish between the 'attributes of space' and the 'spatial attributes'. The former are parameters relative to the quality of a given environment or room, for example, reverberance or warmth [24]. On the other hand, a spatial attribute can be defined as "the three-dimensional characteristic of the components of a spatial audio scene" [25]. This means that the listener is involved. The increase of spatial sound technologies as 5.1 reproduction, ambisonic, binaural, WFS, tends to develop new approaches about the quality which is the spatial quality. Consequently, the perception of sound is modified and gives different sensations to the listeners. Some attributes have been highlighted by recent studies as localisation, immersion, depth, width, distance, spaciousness... (See Table 1)

1.4 Meaning and weight of attributes

Hedonic judgements are prone to biases in listening tests [26]. Hedonic attributes (e.g., "annoying", "pleasant" [13]) refer to the likes, dislikes or preferences of a given person [1]. By contrast, in sensory judgements, listeners are asked to assess sound character, such as timbre, envelopment etc.

In a test protocol a huge problem is the definition of each attribute. Indeed, the understanding of the features by the different assessors must be alike to not bias the results. In many experiments, the definitions of attributes are peculiar to the test. Let us consider several definitions of attributes by various authors as examples:

Clarity:

- The clearer the sound, the more details you can perceive in it. Choose the sound that appears clearer to you. [27]
- This attributes describes if the sound appears clear or muffled [21]

Naturalness:

- A sound is natural if it gives you a realistic impression, as opposed to sounding artificial [27].

Nakayama & al. [20]	Gabrielsson & Sjögren [16]	Toole [17]	Berg & Rumsey [18]
Sensation of clearness	Clearness / Distinctness	Clarity definition	Localisation
	Brightness – Darkness	Brightness	Preference
	Sharpness/hardness – Softness	Softness	Envelopment
Sensation of fullness	Fullness – Thinness	Fullness	Width
	Feeling of space	Pleasantness	Presence
	Disturbing sounds	Hiss, noise, distortion	Naturalness
Depth of the image	Nearness	Impression of distance/depth	Source distance
	Loudness	Definition of sound image	Source width
		Continuity of the sound stage	background noise level
		Fidelity	
		Abnormal effects	
		Reproduction of ambiance, spaciousness & reverberation	
		Perspective	
		Overall spatial rating	

Koivuniemi & Zacharov [19]	Guastavino & katz [13]	Lorho [21]	Choisel & Wickelmaier [6]
Tone Color	Coloration	Clarity	Clarity
Richness	Presence	Richness	Brightness
Hardness	Readability	Sense of distance	Spaciousness
Emphasis	Stability	Sense of direction	Envelopment
Naturalness	Naturalness/ Realism	Sense of movement	Naturalness
Sense of direction	Distance	Ratio of localisation	Elevation
Sense of depth	Localization	Quality of echo	Width
Sense of space	Spatial distribution of sound	Sense of space	Distance
Sense of movement	Spectral balance	Amount of echo	
Penetration		Balance of space	
Distance to events		Separability	
Broadness		Broadness	
		Distorsion	
		Disruption	
		Tone color	
		Balance of Sounds	

Table 1: Synthesis of several studies on quality attributes [22][23]

- Naturalness describes how well the perceived events conform to what the subjects consider as realism [19]
- How similar to a natural (i.e. not reproduced through e.g. loudspeakers) listening experience the sound as a whole sounds. [28]

Envelopment:

- A sound is enveloping when it wraps around you. A very enveloping sound will give you the impression of being immersed in it, while a non-enveloping one will give you the impression of being outside of it. [27]
- The extent to which the sound source envelops/surrounds/exists around you. The feeling of being surrounded by the sound source [28]

Richness:

- This describes the homogeneity of the timbre of a sample [19]
- This attribute describes how rich and nuanced the audio sample is overall, and relates to a combination of harmonics and dynamics perceived in the sample [21]

Sometimes, definitions seem alike, but listeners may interpret them differently. For example, a definition associates the richness with timbre, and another one uses the terms: nuances and dynamics.

A question has arisen about the influence of timbral attributes, and independently spatial attributes, on the overall quality. For audio codec rating, a thin distinction can be made about the language, the most suitable term

is fidelity and not quality because there is a comparison with a reference. The basic audio quality (BAQ) is more correlated to the timbral fidelity than to the spatial fidelity though the latter has a significant relevance [29]. Equation (1) allows one to quantify this contribution in the overall quality [30].

$$BAQ = 0.80\ Timbral + 0.30\ Frontal + 0.09\ Surround - 18.7 \quad (1)$$

2 METHODOLOGY

A general procedure can be explained to create a new test methodology. Let us describe a three step sensory evaluation process [31]: (See Fig. 2)

. Evoke: regroups the panel composition, the stimuli set, the evaluation parameters and scale.

. Measure: the scoring.

. Analyze and interpret: choice of analysis method (statistical or linguistic) and how results are interpreted.

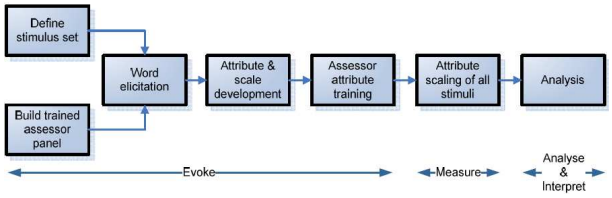


Figure 2: descriptive analysis process [23]

The listening conditions must not be omitted and should be defined accurately. For multichannel sound, the listening configuration is specified in the ITU-R BS 775.1 recommendation [32]. Figure 3 illustrates a room configuration for 5.1 multichannel reproductions.

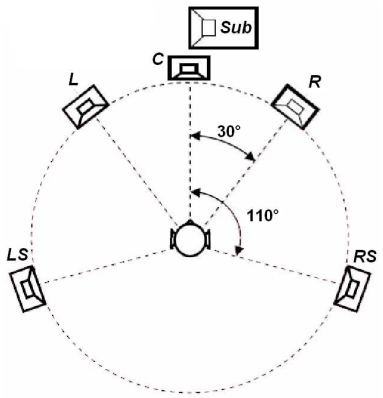


Figure 3: 5.1 room configuration [32]

In the sound field, it is a great benefit to have an agreed approach established by experts [10]. Two standards recommended by the International Telecommunication Union (ITU) are available for the assessment of the audio quality. ITU-R BS.1116 (Triple stimuli with hidden reference) [33] is used for small impairments (high audio quality), whereas the MUSHRA method ("MUlti Stimuli test with Hidden Reference and anchor", ITU-R BS.1534) [34] is employed to evaluate intermediate qualities. Both are mainly used in audio codecs assessment despite the availability of other test methods e.g. A/B-test, ITU-T P.910 (Absolute Category Rating), ITU-R BT.500 or ITU-T P.800 employed in speech or videos evaluation. One of their drawbacks is that they are restricted to the evaluation of the overall audio quality. Consequently, the assessors must estimate the audio excerpts along a single axis of judgment.

2.1 The ITU BS 1116

The ITU-R BS 1116 is a methodology used for the evaluation of small impairments in audio contents. This test is based on triple stimuli with hidden reference [33]. A panel of twenty expert listeners is often sufficient to get reliable results. Experts mean that, they have expertise in the detection of small impairments. The scores by individual assessor can generate biases in results. Thus, for a fair evaluation, the subjects that prove to be less critical or too critical can be rejected. Another solution is to increase the number of subjects.
The duration of the sequences is between 10 and 25 seconds. During the test, the assessors are asked by trial to evaluate three stimuli A, B, C where 'A' is the reference. 'B' and 'C' are randomly the hidden reference and the degraded item. Assessors can listen to the audio excerpts as many times as they wish. The degradation scale is continuous with five grades between 1 (very annoying) and 5 (imperceptible), and a resolution at one decimal. For each trial, when the hidden reference is identified, the altered version is scored along the degradation scale. (See Fig. 4). For the statistical analysis, an ANOVA model is often applied as the first stage.

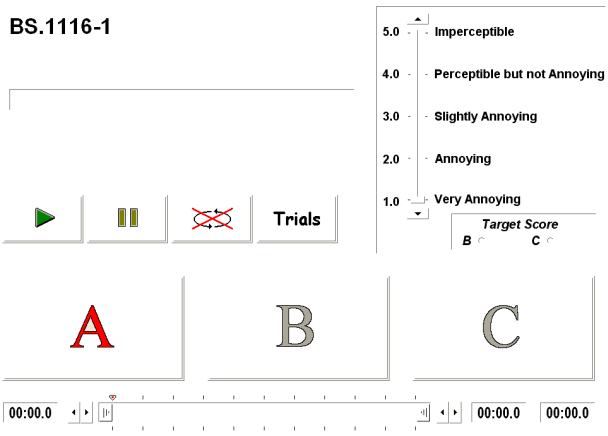


Figure 4 : Example of ITU-R BS.1116-1 Interface

2.2 Mushra

The ITU-R BS 1534-1 is usually called MUSHRA for "MUlti Stimulus test with Hidden Reference and

Anchor" [34]. MUSHRA has been acknowledged as more appropriate for the assessment of intermediate audio quality [35] with no accompanying pictures. MUSHRA and ITU-R BS 1116 have several common features such as the panel selection. Furthermore, there is a reference which is the unprocessed signal (full bandwidth).
The proposed audio excerpts should not be longer than 20s to prevent fatigue by listeners and to reduce the listening test duration
In a trial, many items are presented at the same time. The hidden reference and, at least, one anchor the low-pass filtered of the reference at 3.5 kHz are among the various versions of a sound excerpt. Other anchors can be added as bandwidth limitation of 7 or 10 kHz, reduced stereo image, additional noise…
The aim of the method is to identify the hidden reference and to score all items on a continuous quality scale labeled from 0 to 100, with five categories (bad, poor, good, fair, excellent) (Fig. 5). The hidden reference is scored 100.

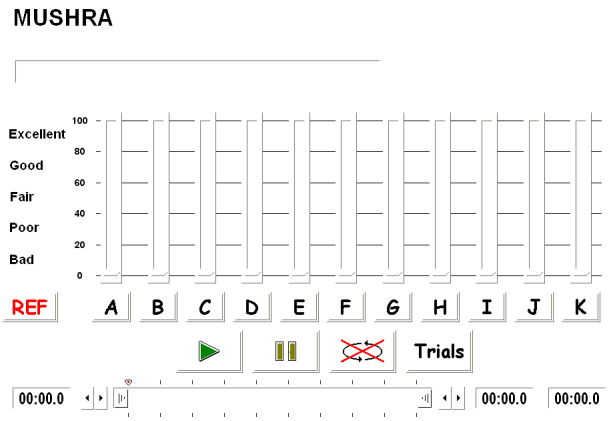


Figure 5 : Example of MUSHRA Interface

Recent studies highlighted biases in MUSHRA test. For example, "the scores may be shifted upwards if additional low quality items are included in the test" or "for the same range of stimuli the MUSHRA test can yield different result depending on the distribution of the stimuli" [36].

The main drawback of ITU methods is that their assessment is about only the overall quality of audio contents. Through two attributes 'front image quality' and 'Impression of surround quality' are proposed in the recommendations, the test value remains limited.

2.3 Other subjective methods

In 1997, the European Broadcasting Union (EBU) proposed the tech 3286-E [37]. This test protocol takes into account quality attributes and was developed for the evaluation of 'classical music' material. However, it can be applied to a live acoustical performance within a real space. This method includes seven specific parameters for the multichannel assessment [37].
Front image quality: the front sound images appear to have the correct and appropriate directional distribution.
Side and rear sound quality: the side and rear sounds appear to have the correct and appropriate balance.
Spatial impression: the performance appears to take place in an appropriate space.
Transparency: the details of the performance can be clearly perceived.
Balance: the individual sound sources and the ambience appear to be properly balanced in the general sound image.
Sound colour: the accurate representation of the characteristic sound of the sources.
Freedom from noise and distortions: absence of various perceptible disturbances.
Main impression: a subjectively weighted average of the other parameters.
Some sub-parameters are proposed for each attribute. Let us cite, location accuracy, spatial reality, reverberation, envelopment, intelligibility, timbre, noise, distortion.
The scale is non continuous and consists of six ranking categories (see Table 2)

Grade	Quality	Impression
1	**Bad**	Very annoying defects
2	**Poor**	Too many annoying defects
3	**Fair**	A number of annoying defects
4	**Good**	Some slightly annoying defects
5	**Very good**	Some perceptible but not annoying defects
6	**Excellent**	No perceptible defects

Table 2 : Quality ranks scale of EBU 3286 [37]

The test requires an important training step for a good definition of these attributes. There is no reference. The panel of listeners listens to each excerpt, checks the evaluation grade and notes down the verbal comments about each of the seven attributes. A discussion session follows the evaluations to permit the listeners to exchange theirs views. The results have to be analyzed with non-parametric methods. Their graphical representation is a radar chart (See Fig. 6).

The ITU-R BS 1284 gathers general methods for the subjective assessment of sound quality [38]. It proposes three scales, degradation and quality scales (see the paragraph hereabove) together with a comparison scale with seven grades (-3 much worse, -2 worse, -1 slightly worse, 0 the sale, 1 slightly better, 2 better, 3 much better). test can consist in a single presentation, or a paired comparison of items where one of them can be

paired comparison of items where one of them can be the reference, or a multiple comparison of stimuli with or without reference. This method includes many parameters detailed in the EBU 3286.

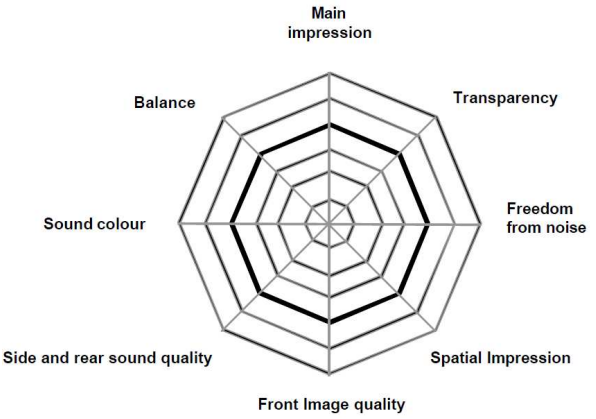


Figure 6: Radar chart from EBU 3286 – S1 [37]

2.4 Objective methods

The cost in time and money of a subjective evaluation explains why objective methods have been developed for quality assessment.

For speech assessment, the model is PESQ (Perceived Evaluation Speech Quality). It is a psychoacoustic intrusive model; the forecast of quality is based on the comparison between referring signal and altered signal.

In audio field, the model of reference is PEAQ (Perceptual Evaluation of Audio Quality) [39]. It is an artificial ear-like tool. PEAQ is a standardized algorithm developed to evaluate the perceived audio quality. It is based on psycho-physic principles, notably on masking and on cognitive effects. This model is declined in two versions 'basic' and 'advanced'. The ear model of the basic version of PEAQ uses only the fast Fourier transform (FFT) and is employed for applications that require high processing speed. The advanced version is heavier to run but the calculus is more accurate. It based on FFT as well as on the filter bank ear model. Both versions produce many model output variables (MOVs) such as envelope modulation, detection probability or noise-to-mask ratio. For the final mapping, the basic version uses eleven MOVs whereas the advanced version uses five MOVs [40]. The cognitive model is a neural network that uses the MOVs calculated previously. At the end of the network, the software gives, per sequence, a value for overall quality called objective difference grade. The validation of the objective methods is based on the correlation between their ODGs and the corresponding subjective difference grade (SDG) resulting from dedicated subjective tests.

3 FURTHER STUDIES

Future investigations consist in running tests, so as to find a method dedicated to the assessment of surround sounds. New scales have to be tested. Main attributes could be included in the listening test procedure. A new interface will need to be developed in order to fulfill the found features of this future works.

4 CONCLUSIONS

This paper is a review of the spatial audio attributes. It recalls the different methods currently used to evaluate the quality of multichannel sounds. It shows that many features specific to spatial audio are well defined and can be included in a new test methodology.

REFERENCES

[1] S. Bech, "Methods for subjective evaluation of spatial characteristics of sound", *Proceedings of the 16th AES International Conference on Spatial Sound Reproduction*, Audio Eng. Soc. (1999).

[2] J. Berg and F. Rumsey. "Spatial attribute identification and scaling by repertory grid technique and other methods", Proceedings of *the 16th AES International Conference on Spatial Sound Reproduction*, Audio Eng. Soc. (1999)

[3] G. Kelly, *The psychology of personal constructs*. Norton, New York, USA. (1955)

[4] J. Berg. "Evaluation of perceived spatial audio quality". Proceedings of *the 9th World Multi-Conference on Systemics, Cybernetics and Informatics* vol. 4, no. 2, pp.10-14. (2005)

[5] J. Berg. "OPAQUE – A tool for the elicitation and grading of audio quality attributes". Presented at *AES 118th Convention*, Audio Eng. Soc. (2005)

[6] S. Choisel and F. Wickelmaier. "Extraction of auditory features and elicitation of attributes for the assessment of multichannel reproduced sound". Presented at *AES 118th Convention*, Audio Eng. Soc. (2005)

[7] B. Ganter and R. Wille, *Format Concept Analysis: Mathematical Foundations*, Springer, Berlin. (1999)

[8] J.P. Doignon and J.C. Falmagne, *Knowledge Spaces*, Springer, Berlin. (1999)

[9] J. Heller, "Representation and Assessment of Individual Semantic Knowledge", *Methods of*

Psychological Research, vol 5, n°2, pp. 1-37 (2000)

[10] S. Bech and N. Zacharov, *Perceptual Audio Evaluation Theory, Method and Application*. Wiley, (2006)

[11] F. Rumsey. "Subjective assessment of the spatial attributes of reproduced sound" *Proceedings of the 15th AES International Conference on Audio, Acoustics & Small Spaces*, Audio Eng. Soc. (1998).

[12] P. Susini, S. McAdams, S. Winsberg. "A multidimensional technique for sound quality assessment". *Acustica* vol. 85, pp.650-656. (1999)

[13] C. Guastavino and B. Katz. "Perceptual evaluation of multi-dimensional spatial audio reproduction". *Journal of the Acoustical Society of America* vol. 116, no. 2, pp. 1105–1115. (2004)

[14] S. Le Bagousse. "Etude perceptive et acoustique du timbre de la voix chantée dans le contexte des répertoires de tradition orale" Master thesis, Université de Paris 6, Ircam. (2007)

[15] J.M. Grey, "Multidimensional perceptual scaling of music timbres". *Journal of the Acoustical Society of America* vol. 61 pp. 122-135. (1977)

[16] A. Gabrielsson and H. Sjögren, "Perceived sound quality of sound reproduction systems", *Journal of the Acoustical Society of America* vol. 65, no. 4. (1979)

[17] F. Toole. "Subjective measurements of loudspeaker sound quality and listener performance". *Journal of Audio Engineering Society* vol. 33, pp. 2-32. (1985)

[18] F. Rumsey and J. Berg, "Verification and correlation of attributes used for describing the spatial quality of reproduced sound", Presented at the *AES 19th International Conference on Surround Sound*. Audio Eng. Soc. (2001)

[19] K. Koivuniemi and N. Zacharov, "Unraveling the perception of spatial sound reproduction: Language development, verbal protocol analysis and listener training", *Proceedings of the AES 111th Convention*, Audio Eng. Soc. (2001)

[20] T. Nakayama, T. Miura, O. Kosaka, M. Okamoto and T. Shiga. "Subjective assessment of multichannel reproduction". *Journal of Audio Engineering Society* vol. 19, no. 9, pp. 744-751. (1971)

[21] G. Lorho. "Evaluation of spatial enhancement systems for stereo headphone reproduction by preference and attribute rating". Presented at *AES 118th Convention*. Audio Eng. Soc. (2005)

[22] J. Berg and F. Rumsey, "Systematic evaluation of perceived spatial quality", Presented at the *AES 24th International Conference on Multichannel Audio*. Audio Eng. Soc. (2003)

[23] T.H. Pedersen and N. Zacharov. "How many psycho-acoustic attributes are needed?" *Acoustics'08 Paris*. (2008)

[24] F. Rumsey. "Spatial quality evaluation for reproduced sound: terminology, meaning and a scene-based paradigm". *Journal of Audio Engineering Society* vol. 50, no. 9, pp. 651-666. (2002)

[25] F. Rumsey. "Spatial audio and sensory evaluation techniques – context, history and aims". In Proceedings of *Spatial audio & sensory evaluation techniques conference*, Guilford, UK (2006)

[26] S. Zielinski. "On some biases encountered in modern listening tests". In Proceedings of *Spatial audio & sensory evaluation techniques conference*, Guilford, UK (2006)

[27] S. Choisel and F. Wickelmaier. "Evaluation of multichannel reproduced sound: Scaling auditory attributes underlying listener preference". *Journal of the Acoustical Society of America* vol. 121, no. 1, pp. 388-400. (2007)

[28] J. Berg and F. Rumsey. "Validity of selected spatial attributes in the evaluation of 5-channel microphone techniques". Presented at *AES 112th Convention*. Audio Eng. Soc. (2002)

[29] P. Marins, F. Rumsey and S. Zielinski. "Unraveling the relationship between basic audio quality and fidelity attributes in low bit-rate multi-channel audio codecs". Presented at *AES 118th Convention*. Audio Eng. Soc. (2008)

[30] F. Rumsey, S. Zielinski, R. Kassier and S. Bech. "On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality". *Journal of the*

Acoustical Society of America vol. 118, no. 2, pp. 968-976. (2005)

[31] H.T. Lawless and H. Heymann. *Sensory evaluation of food*, Chapman and Hall, (1998)

[32] ITU-R, Recommendation BS.775-1. Multichannel stereophonic sound system with and without accompanying, International Telecommunications Union Radio-communication Assembly. (1992)

[33] ITU-R, Recommendation BS.1116-1. Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems, International Telecommunications Union Radio-communication Assembly. (1997)

[34] ITU-R, Recommendation BS.1534-1. Method for the subjective assessment of intermediate quality level of coding systems, International Telecommunications Union Radio-communication Assembly. (2003)

[35] G.A. Soulodre and M.C. Lavoie, "Subjective evaluation of large and small impairments in audio codecs". Proceedings of *the 17th AES International Conference on High-quality Audio Coding*, Audio Eng. Soc. (1999).

[36] S. Zielinski, P. Hardisty, C. Hummersone and F. Rumsey. "Potential biases in MUSHRA listening tests". Presented at *AES 123th Convention*. Audio Eng. Soc. (2007)

[37] EBU. Assessment methods for the subjective evaluation of the quality of sound programme material – Music. (1997)

[38] ITU-R, Recommendation BS.1284-1. General methods for the subjective assessment of sound quality, International Telecommunications Union Radio-communication Assembly. (1997-2003)

[39] ITU-R, Recommendation BS.1387-1. Methods for objective measurements of perceived audio quality, International Telecommunications Union Radio-communication Assembly. (1998)

[40] T. Thiede, W.C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J.G. Beerends, C. Colomes, M. Keyhl, G. Stoll, K. Brandenburg and B. Feiten. "PEAQ – The ITU standard for objective measurement of perceived audio". *Journal of Audio Engineering Society* vol. 48, no. 1, pp. 3-21. (2000)

Audio Engineering Society

Convention Paper

Presented at the 129th Convention
2010 November 4–7 San Francisco, CA, USA

Families of Sound Attributes for Assessment of Spatial Audio

Sarah Le Bagousse[1], Mathieu Paquier[2], and Catherine Colomes[1]

[1] Orange Labs France Télécom R&D, Cesson Sévigné, 35510, France
sarah.lebagousse@orange-ftgroup.com

[2] Laboratoire d'Informatique des Systèmes Complexes (LISyC), Brest, 29000, France
Mathieu.Paquier@univ-brest.fr

ABSTRACT

Over the last years, studies have highlighted many features liable to be used for the characterization of sounds by several elicitation methods. These various experiments have resulted in the production of a long list of sound attributes. But, as their respective meaning and weight are not alike for assessors and listeners, the analysis of the results of a listening test based on sound criteria remains complex and difficult. The experiments reported in this paper were aimed at shortening the list of attributes by clustering them in sound families from the results of two semantic tests based on either a free categorization (i) or use of a multi-dimensional scaling method (ii).

1. INTRODUCTION

Spatial audio is, nowadays, among the main subjects of focus for the broadcasting of audio contents. The constant development of spatial technologies is at the origin of the use of new terms to qualify sounds. This is why spatial attributes of sound were deeply studied over the last few years. Nevertheless, other types of attributes such as timbral ones are still evaluated. The sound attributes have been highlighted by various direct or indirect elicitation methods such as, for example, the repertory grid technique [1], the perceptual structure analysis [2], multidimensional scaling [3]. Hence, a list of sound attributes has been established and includes, for example, coloration, envelopment, noise, timbre, brightness, localization, hiss, distortion, volume, richness… But, because of the lack of clear definition of these terms in the literature, the way they are understood is dependent upon the individuals. Furthermore, as the meanings of certain criteria seem to be very close, this causes biases in listening tests. Though it would be worth including the main sound attributes in the listening tests currently proposed, it appears that, in most of sound assessment methods such as ITU

standards, the evaluation is focused on only a single axis called basic audio quality (BAQ). The BS.1116 [4] or MUSHRA [5] test methodologies both take into account the front image quality and impression of surround quality besides BAQ, but this remains limited. Including attributes in an audio test would permit an enhancement of the accuracy of the results and, for the audio coding assessment, give some clues about the sound characteristics that have been impaired. Introducing sound families in a listening test should be easier than proposing numerous attributes to the listeners. A sound family is composed of a group of attributes. According to the literature, some studies have dealt with two main families of attributes, *i.e.* the timbral ones and the spatial ones, but no real experiments had been carried out to precisely define them. They have led to the development of the Mural model (MUlti-level auditoRy Assessment Language) by Letowski where both groups of attributes and their subsets are defined [6]. Berg and Rumsey used three kinds of sound features (timbral, spatial and technique) to classify attributes [7], but the information about the origin of this classification are missing. (See figure 1)

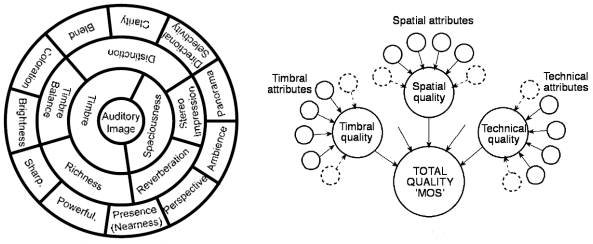


Figure 1: Mural [6] and sound categories [7]

In the investigations reported here, two semantic experiments were conducted in parallel so as to highlight several families of sound attributes. At first, a list of attributes selected among all of the features elicited by various studies on the audio assessment [8] was established and presented to the assessors. The first experiment consisted in a linguistic test aimed at establishing a free categorization of attributes: the assessors were, indeed, asked to classify the attributes from the list into different groups. The second experiment was based on the use of a multidimensional scaling method (MDS). The assessors were asked to score the perceived similarity between two terms on a dissimilarity scale. Then, the results from both experiments were compared to validate the sound families.

2. SOUND ATTRIBUTES

In order to avoid biases in both tests through the classification of antonym terms into two different categories, some of them were excluded from the list of attributes proposed to the assessors. For example, *softness* was deleted conversely to *hardness.* The features liable to be ambiguous, e.g. *volume,* which can refer to either loudness or auditory size, were also excluded. At last, the final list was composed of 28 attributes that represent the various aspects of sound characteristics (Table 1).

Sound Attributes		
Background noise	Equalization	Realism
Brightness	Fidelity	Reverberation
Clarity	Hardness	Richness
Coloration	Hiss	Sharpness
Depth	Homogeneity	Spatial distribution
Disruption	Hum	Spatialization
Distance	Immersion	Stability
Distortion	Localization	Tone Color
Dynamics	Noise	Width
Envelopment		

Table 1: List of attributes

The whole procedure was made in French, but for a better readability of this paper, they are translated into English.

3. TEST CONDITIONS

The assessor population consisted of 18 people considered as experts because of their solid background acquired by experience through audio testing or work in audio or musical field. The assessor population was split into two equal groups, and then the first half was submitted to the free categorization test while the second one was given the MSD method-based test. Each group made the second test a week later.

4. FREE CATEGORIZATION

4.1. Test procedure

The list of 28 attributes was submitted to the assessors for free categorization. The meaning of each attribute was supposed as known. When it was not the case, none of the assessors was allowed to look for the definition of

a given attribute. They had to refer to their own interpretation. They were asked to group the 28 attributes into, at least, 2 families and, at the most, 5 ones. They had also to find a name for each of the constituted sound families. They were allowed to explain their choices and leave comments.
A specific interface was developed to permit the use of a drag and drop to constitute sound families (Fig. 2)

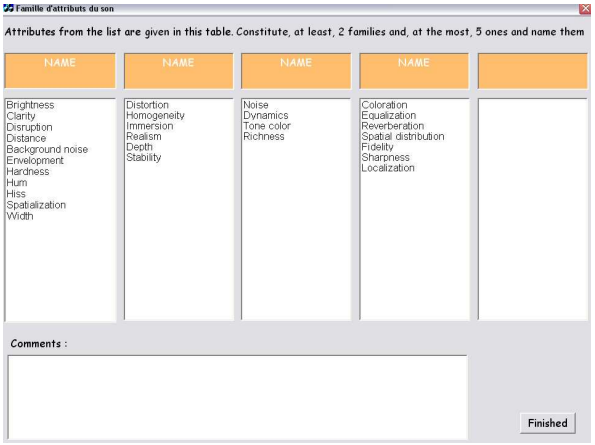


Figure 2: Interface of the free categorization test

4.2. Results

Eighteen tables containing the categories made by each assessor were analyzed. The total number of families of attributes was 73. Nine assessors classified the proposed attributes into 4 families, 5 others made 5 families and 4 other ones constituted 3 families. A matrix (28*73) of all attributes and families was analyzed by clustering with the Clustergram function included in the Matlab Statistics Toolbox. This function realizes an ascending or agglomerative hierarchical clustering. Each point is successively merged by the closest cluster.

The distance between two clusters can be analyzed by different ways. Here, the Ward method was chosen as agglomerative technique. The main difference with other approaches is the analysis of variance instead of distances. The number of clusters is reduced to minimize the loss (the error sum of squares) between each group [9]. The Ward linkage operates the within-cluster sum of squares defined as the sum of the squares of the distances between all of the objects within a cluster and the cluster centroid.

$$d^2(r,s) = n_r n_s \frac{\left\|\bar{x}_r - \bar{x}_s\right\|_2^2}{n_r + n_s} \quad \text{with} \quad \bar{x}_r = \frac{1}{n_r}\sum_{i=1}^{n} x_{ri}$$

where $\|\ \|_2$ is the Euclidean distance and $\bar{x}_r$ and $\bar{x}_s$ are the centroids of the clusters, *r* and *s*.

A dendrogram displays the groups of attributes (see Annexe 1). It allowed us to clearly identify 3 families.
However one of those families is divided into 2 clusters.
This figure enables one to associate the name given by assessors to the attribute cluster. All these names can be associated by synonymous, lexical devices, stemming, and lemmatization in order to name the families. (Annexe 2)

5. MULTIDIMENSIONAL SCALING

5.1. Test procedure

To define the sound families, a second test based on an indirect elicitation method, i.e. the *MultiDimensional Scaling* (MDS) [10] was proposed. It was aimed at revealing various dimensions and 'hidden meaning' in the data [11]. It is based on the dissimilarity between a pair of items.

Two attributes from the previous list of 28 attributes are presented to the assessors in order they assess the similarity between them. The proposed scale is a dissimilarity continuous scale from 0 to 1. It was proved that semantic intervals between labels of a scale are unequal [12]. Thus for this test, the dissimilitary scale is without label except at the endpoints called respectively very different and very similar. The perceived similarity or dissimilarity is interpreted as a measure of the distance between various terms.
At first, the assessors were submitted to the test for 5 minutes as training in order to familiarize them with the scoring method and the terms to be compared.
Then, all of the pairs were presented to the assessors in a random order. For n stimuli, the number of pairs is *n*(n-1)/2*, and thus there are 28*27/2=378 combinations. The completion of the full test took one hour.

A Matlab interface was devised for this experiment (Fig. 3). Each assessor was asked to score the perceived similarity between 2 attributes by moving the cursor to the desired position along the slider. When the assessor was satisfied with his/her own grading, he/she pressed the "Next" button. The score was backed up and the

next pair was displayed. At the end of the process, the symmetric dissimilarity matrix was written.

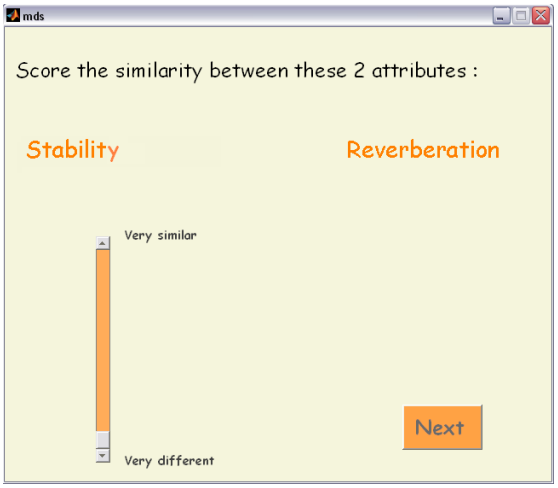


Figure 3: MDS test interface

5.2. Results

Indscal (Individual Differences Scaling) [13] algorithm was chosen as a MDS non metric method for data treatment because it takes into account the inter-individual difference. Each dimension is weighted among subjects. SPSS 'Statistical Package for Social Science' software realizes the analysis of the dissimilarity matrix. This application enables one to develop solutions from 2 up to 6 dimensions.

The results produced by each assessor were stored in a symmetric dissimilarity matrix, further used to create multidimensional space with the attributes.

The output parameters express the degree of fit between the build space and the original data (set of dissimilarities). Stress is the difference between distance in the MDS space and the distance that best fits the dissimilarity and RSQ is squared correlation (proportion of variance) of the distances with the mismatches [14]. These measures help to determine the number of dimensions (Fig. 4)

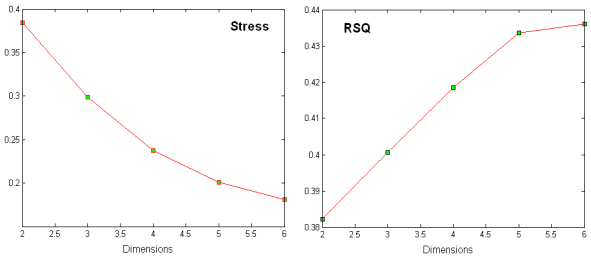


Figure 4: Stress and RSQ values.

28 attrib. / 18 asses	2 dim.	3 dim	4 dim	**5 dim**	6 dim
Stress	0.03851	0.29884	0.23757	**0.20091**	0.18118
RSQ	0.38228	0.40062	0.41849	**0.43363**	0.43599

Table 2: Stress and RSQ for each dimension

Table 2 shows that the RSQ values for dimension 5 and dimension 6 are close. It is worth noting the break on the straight line (Fig. 4), which indicates the lack of enhancement of the RSQ value by a 6th dimension. Thought it is less obvious, the same analysis is suitable for the Stress value. The most suitable number of dimensions is 5 for the perceptive space (stress = 0.20, RSQ = 0.43).

Weights of dimension 1 and dimension 2 are higher than those of other dimensions:

dim1: 0,1550 dim2: 0.1040 dim3: 0.0699
dim4: 0.0530 dim5: 0.0517

The main objective of this experiment was to highlight the resulting clusters and therefore the association between attributes. The identification of dimensions was of less interest.

Figure 5 shows the perceptual space obtained by projection of the attributes upon dimension 1 and dimension 2 considering their coordinates in the space with 5 dimensions.

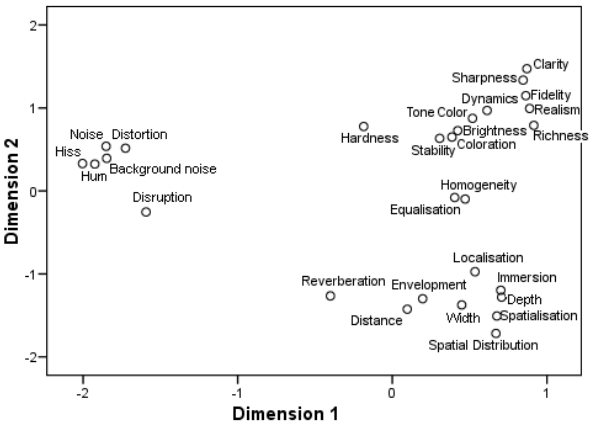


Figure 5: Perceptual space (dimensions 1 and 2)

Table 3 gives the three groups clearly observed on this 2-dimension representation.

Group 1	Group 2	Group 3
Background noise	Fidelity	Reverberation
Noise	Hardness	Spatialization
Distortion	Richness	Spatial distribution
Disruption	Homogeneity	Localization
Hiss	Sharpness	Width
Hum	Tone Color	Distance
	Coloration	Envelopment
	Brightness	Depth
	Clarity	Immersion
	Dynamics	
	Realism	
	Stability	
	Equalization	

Table 3: Attribute groups on the 2-dimension representation.

One should note that homogeneity and equalization seem to be apart. But a projection upon dimensions 2 and 3 showed the split of group 2 into 2 categories: i) richness, hardness, tone color, coloration, clarity and equalization and ii) realism, fidelity, stability, sharpness, homogeneity and dynamics (Fig. 6).

Extracting relevant information from dimensions 4 and 5 proved to be difficult and did not allow us to refine the families. Thus 3 sound families were extracted from the MDS experiment.

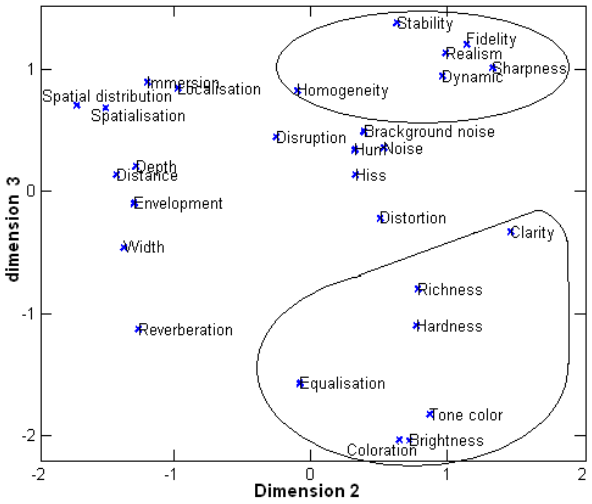


Figure 6: Dimension 2 and dimension 3

6. DISCUSSION AND CONCLUSION

6.1. Sound Families

The experiments carried out in this study permitted us to highlight 3 families of sound attributes as follows:

- **Defaults**: are interfering elements or nuisances present in a sound, e.g. noise, distortion, background noise, hum, hiss, disruption …

- **Space**: refers to spatial impression-related characteristics, e.g. depth, width, localization, spatial distribution, reverberation, spatialization, distance, envelopment, immersion …

- **Timbre**: this family is split into 2 subfamilies : the first one deals with the sound color, e.g. brightness, tone color, coloration, clarity, hardness, equalization, richness…
  The second one composed of homogeneity, stability, sharpness, realism, fidelity and dynamics describes the timbre but can be also related to other characteristics of sound.

6.2. Conclusion

From free categorization and MultiDimensional Scaling-based tests, this experimental study highlighted 3 sound families. These results are correlated to the previous studies which dealt with clusterization of sound attributes. It permitted us to clarify the meanings of sound families and to make rid of the biases currently

created in listening tests by certain definitions of sound attributes. It would be, therefore, worth including these sound families in listening tests aimed at evaluating audio quality. Such tests could be improved by asking assessors to judge along several axes further used which will be the established sound families. The interest is to verify if one of the families is more involve in the audio quality assessment.

7. REFERENCES

[1] J. Berg and F. Rumsey. "Spatial attribute identification and scaling by repertory grid technique and other methods", presented at *the 16th AES International Conference on Spatial Sound Reproduction*, Finland, 1999 April 10-12.

[2] S. Choisel and F. Wickelmaier. "Extraction of auditory features and elicitation of attributes for the assessment of multichannel reproduced sound". Presented at *AES 118th Convention*, Barcelona, Spain, 2005 May 28-31.

[3] J.M. Grey, "Multidimensional perceptual scaling of music timbres". *Journal of the Acoustical Society of America*, vol. 61, pp. 122-135 (1977 May).

[4] ITU-R Recommendation BS.1116-1., "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems", International Telecommunications Union, Radio-communication Assembly. (1997)

[5] ITU-R, Recommendation BS.1534-1., "Method for the subjective assessment of intermediate quality level of coding systems", International Telecommunications Union, Radio-communication Assembly. (2003)

[6] T. Letowski, "Sound Quality Assessment: Concepts and Criteria", presented at the *AES 87th Convention*, New York, USA, 1989 October 18-21.

[7] J. Berg and F. Rumsey, "Systematic evaluation of perceived spatial quality", presented at the *AES 24th International Conference on Multichannel Audio*, Banff, Canada, 2003 June 26-28.

[8] S. Le Bagousse, C. Colomes and M. Paquier. "State of the art on subjective assessment of spatial audio quality", presented at the *38th AES International Conference on Sound Quality Evaluation*, Pitea, Sweden, 2010 June 13-15.

[9] J.H. Ward, "Hierarchical Grouping to optimize an objective function", *Journal of the American Statistical Association*, vol. 58, pp. 236-244 (1963 March).

[10] S. Bech and N. Zacharov, *Perceptual Audio Evaluation Theory, Method and Application*. Wiley, (2006)

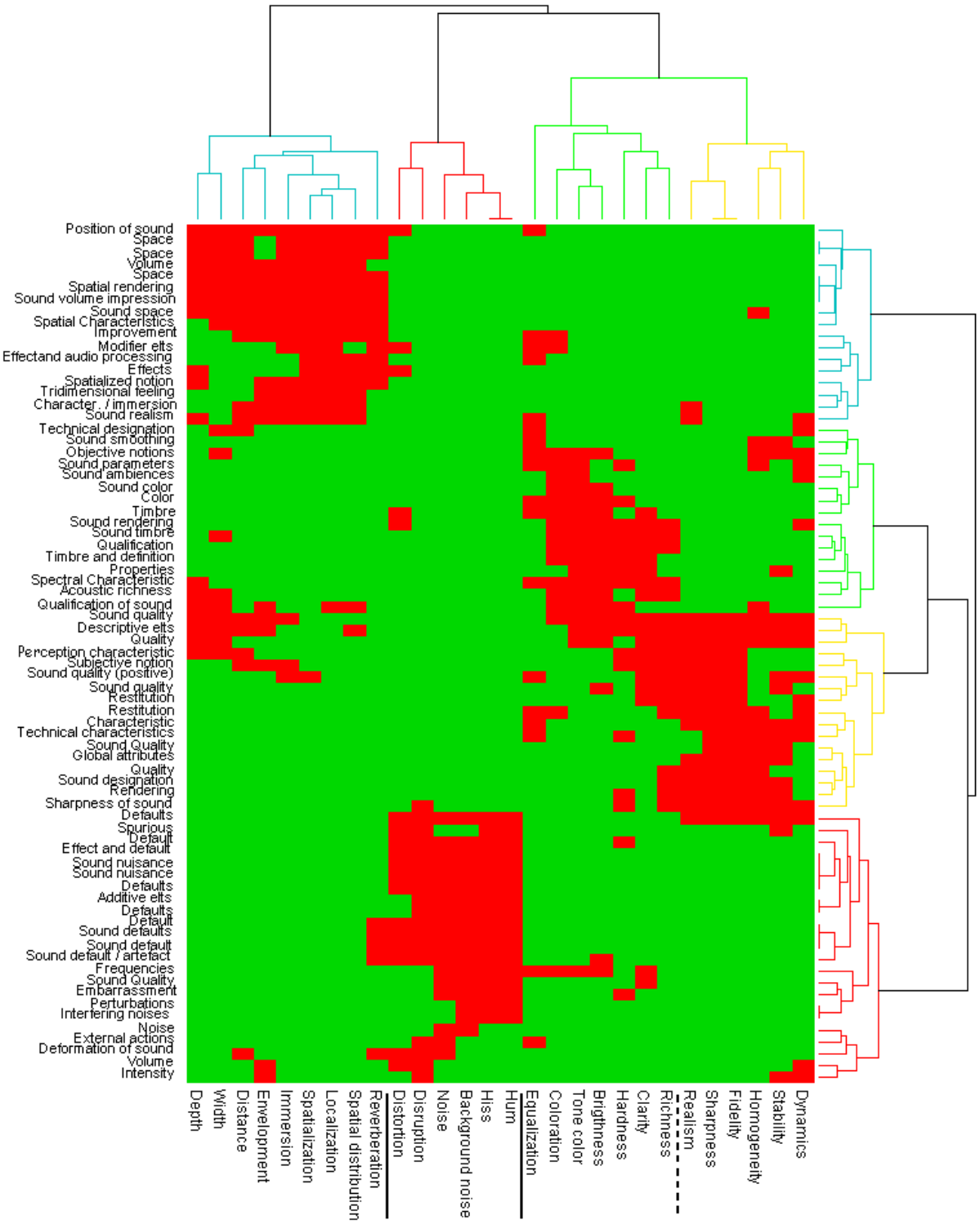
[11] F. Rumsey. "Subjective assessment of the spatial attributes of reproduced sound" presented at *the 15th AES International Conference on Audio, Acoustics & Small Spaces*, Copenhagen, Denmark, 1998 October 31-November 2.

[12] S. Zielinski, P. Brooks and F. Rumsey, "On the use of graphic scales in modern listening tests", presented at the *AES 123th Convention*, New York, USA, 2007 October 5-8.

[13] J.D Carroll and J.J. Chang, "analyses of individual differences in multidimensional scaling via an N-way generalization of 'Eckart-Young' decomposition", *Psychometrika*, vol. 35, n°3, pp. 283-319, (1970 September)

[14] T. Naes and E. Risvik, "*Multivariate analysis of data in sensory science*", Elsevier Science Ldt, (1996)

ANNEXE 1: DENDROGRAM WITH CLUSTERING RESULTS



ANNEXE 2: FAMILIES NAMES FROM CATEGORIZATION TEST

Group 1	Group 2	Group 3	Group 4
Defaults (9 occurrences)	Technical designation	Position of sound	Qualities (6)
Spurious	Sound smoothing	Space (4)	Descriptive elements
Effects	Objective notions	Volume	Characteristics of perception
Sound nuisances	Parameters of sound	Spatial rendering	Subjective notions
Additive elements	Sound ambiences	Impression of volume	Restitution (2)
Artifacts	Color (2)	Spatial characteristics (2)	Characteristics (2)
Frequencies	Timbre (3)	Improvements	Global attributes
Quality of sound	Sound rendering	Modifier elements	Sound designation
Embarrassment	Qualification (2)	Effects and audio processing	Rendering
Perturbations	Properties	Effects	Sharpness of sound
Interfering noises	Spectral characteristics	Spatialized notion	
Noise	Acoustics richness	Tridimensional feeling	
External actions		Realism	
Deformation of sound		Immersion	
Volume			
Intensity			

Audio Engineering Society

Convention Paper

Presented at the 131st Convention
2011 October 20–23 New York, USA

Sound Quality Evaluation based on Attributes - Application to Binaural Contents

Sarah Le Bagousse[1], Mathieu Paquier[2], Catherine Colomes[1] and Samuel Moulin[1]

[1] *Orange Labs, Cesson Sévigné, 35510, France*

[2] *Laboratoire d'Informatique des Systemes Complexes, Brest, 29200, France*

Correspondence should be addressed to Sarah Le Bagousse (sarah.lebagousse@orange-ftgroup.com)

ABSTRACT

The audio quality assessment is based on standards which mainly evaluate the overall quality to the detriment of more accurate sound criteria. On the other hand, a major problem of an assessment based on sound criteria is their meaning and their understanding that have to be the same for each listener. A previous study clustered a list of sound attributes in three main categories called "timbre", "space", "defaults". The work presented here is based on those previous results and aims at tuning a subjective test methodology of spatial audio quality. So the three families were included in a test dedicated to the assessment of spatial audio quality with binaural contents. The test was based on the MUSHRA method but using three anchors specifically to each attribute and without explicit reference. The original version was added as the hidden reference. The aim of the listening test described in this paper was to verify the relevance of those 3 attributes and their influence on the overall quality.

1. INTRODUCTION

Before being broadcasted on services, the quality of audio contents have to be evaluated. Due to the development of spatial technologies, current methods of sound quality assessment reveal some lacks. Indeed standards do not take into account specific features of spatial sound. The basic audio quality (BAQ) is often the only evaluated attribute. According to ITU-R BS.1534 [1], BAQ is the "global attribute used to judge any and all detected differences between the reference and the object". Although it is a sufficient indicator, it would be interesting to obtain some clues on impairments influencing the overall quality. Some attributes, such as coloration, brightness, distortion, localization... have been highlighted by different elicitation meth-

ods. However their definitions and their understandings remain a major problem and it is difficult to include them in a test method [2]. Rather than submitting a list of attributes to listeners, it was possible to gather them in different main sound families. The bias created by specific attributes meanings is therefore reduced. In theirs studies, Rumsey et al [4] [5] defined categories of attributes such as timbral, frontal and surround fidelity attributes. The aim of those experiments was to compare various items to their reference for each one of the 4 fidelity attributes (BAQ, timbral, frontal and surround fidelity). These tests showed that timbral fidelity was more correlated to the BAQ than spatial fidelity. The term fidelity was employed because tests included an explicit reference.
A previous study which aimed at clustering a list of attributes in categories highlighted 3 sound families for qualifying audio contents: "timbre", "space" and "defaults" [3]. The test described in this paper includes these three attributes "timbre", "space" and "defaults" in addition to the overall quality criterium. One of the requirements for the tested method was that there were no explicit reference. Nevertheless, the original version, was included as a hidden reference. Thus the term overall quality was employed instead of BAQ. For the test design, the employed method was inspired by the recommendation ITU-R BS.1534 [1]. The method was applied to binaural contents broadcast on headphones. The aim of the experiment described in this article was to test a quality evaluation method and to prove the influence, precisely the weight, of the attributes timbre, space and defaults on the overall quality in the context of spatial audio assessment.

2. ATTRIBUTES FAMILIES

A previous experiment was run in order to highlight families of sound attributes to describe the quality of spatial audio [3]. Tests consisted in presenting a list of attributes (28) and asking assessors to classify them in some categories. No sound was presented in order to create groups independently of audio restitution systems. Two methods were employed: a multidimensional scaling (MDS) and a clusters analysis. Both tests obtained the same results and thus three families were defined.

- Defaults: are interfering elements or nuisances present in a sound, e.g. noise, distortion, background noise, hum, hiss, disruption...

- Space: refers to spatial impression-related characteristics, e.g. depth, width, localization, spatial distribution, reverberation, spatialization, distance, envelopment, immersion...

- Timbre: this family is split into 2 subfamilies:
  The first one deals with the sound color, e.g. brightness, tone color, coloration, clarity, hardness, equalization, richness...
  The second one composed by homogeneity, stability, sharpness, realism, fidelity and dynamics describes the timbre but can also be related to other characteristics of sound.

3. LISTENING TEST DESIGN

3.1. Panel composition and listening conditions

The panel of assessors consisted in 18 "experts" fulfilling the requirements of MUSHRA method. They are able to detect impairments in audio signals and they have solid musical background due to their job in audio or musical field.

The test took place in a dedicated soundproof booth and was performed on headphone STAX Signature SR-404 (open) and its amplifier SRM-006t.

3.2. Excerpts, anchors and degradations

Seven excerpts were included in this listening test. They were chosen through film, environment and music categories to cover a large range of contents. Three of them were native binaural recordings whereas the 4 others were realized by binaural synthesis of multichannel (5.1) excerpts. A brief description of contents is presented in table 1. Each sequence was twenty seconds long at the maximum according to the recommendation ITU BS.1534 [1].

For each excerpt, eight versions were presented to assessors including the original (unprocessed signal), four codecs and three anchors specific to each quality attribute. These degradations are described in table 2.

"3.5" item was defined as a timbral anchor, "SA" as a spatial anchor and "noise" as a defaults anchor. 3.5 was a low-pass filtered version of the original

Name	Time (s)	Record system	Description
Barber	18	binaural recording	virtual hair cut
Bombarde	22	binaural recording	bombarde player who walked away
Stair	17	binaural recording	people walking down stairs
Marimba	20	synthesis from 5.1	music with marimba
Milanof	20	synthesis from 5.1	music
Star wars	18	synthesis from 5.1	film extract
Tango	18	synthesis from 5.1	music

Table 1: Table of excerpts.

Abrev.	***Kind of degradation***
Original	Original as hidden reference
3.5	3.5 filtered low pass
noise	Pink noise added plus clics
SA	Channel inversion and mono
HEAAC	HE-AAC+ at 40 kbits/s
AMR	Voice AMR at 48 kbits/s
MP3	MP3 at 64 kbits/s
AAC	AAC at 32 kbits/s

Table 2: Table of degradations.

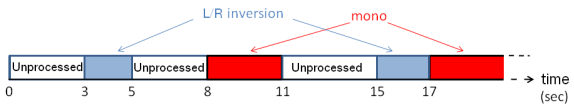


Fig. 1: Spatial anchor description.

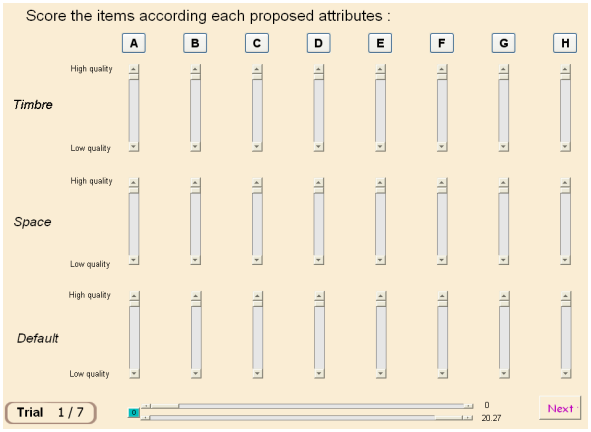


Fig. 2: Interface for three attributes presentation

(at 3.5 kHz). "noise" resulted from the addition of a pink noise and clicks on the unprocessed signal. The spatial anchor "SA" consisted in an inversion between left and right channel, plus few seconds in mono, for more details see figure 1 .

3.3. Test protocol

The test was decomposed in two sessions. The first one was the evaluation of the overall quality and the second one was the assessment of the 3 attributes ("timbre", "space" and "defaults"). The protocol was inspired from Mushra test ITU-R BS.1534 [1]. Stimuli were presented simultaneously in a random order and assessors scored all degradations on a specific quality scale. It was noticed that some biases encountered in subjective methods come from the scale [6] thus the grading scale was without labels except on the end points called "high quality" and "low quality". No number appeared during the grading. Assessors had to place the cursor along the slider range. The test included no explicit reference, though the original version was considered as an hidden reference. Instruction set that the stimuli perceived as the best quality had to be scored at the top of the scale. The interface offered the possibility to reduce the excerpt duration in order to focus on short part of the audio stimuli. Figure 2 show the interface used for the evaluation of the three attributes.

4. RESULTS

The scores and the ranking of processes are presented but the real interest focuses on the evaluation method.

4.1. Overall Quality

The first stage of the test was to evaluate the over-

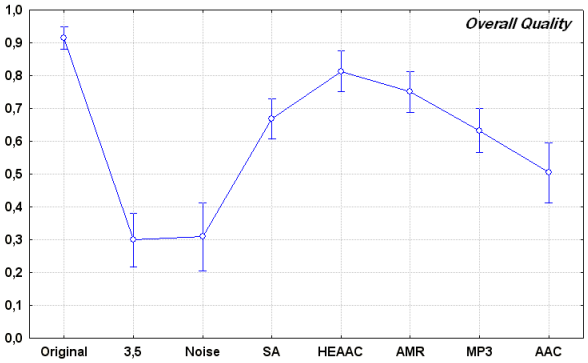


Fig. 3: Means scores and 95% IC of degradations for overall quality.

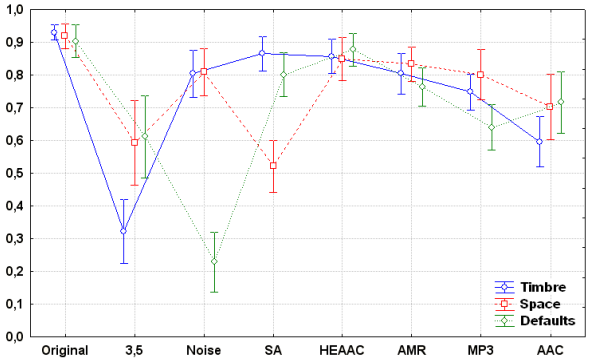


Fig. 4: Means scores and 95% IC of degradations for the three attirbutes timbre, space and defaults.

all quality of the sequences. Results of the assessment are presented in figure 3. The quality scale was converted to a linear numerical scale: 0 for low quality and 1 for high quality. A wide dynamic of the scale was used. The hidden reference was scored 0,92 and was thus identified as the best quality version presented in the test. By contrast, the timbral and defaults anchors obtained the worst scores 0,3. "SA", considered as the spatial anchor, was scored higher than the other anchors at 0.67. Two hypothesis can be drawn from these observations; either spatial degradations slightly affected the overall quality or the proposed anchor was not appropriated as a spatial anchor. Codings were placed in the upper part of the scale but their scores were significantly different.

4.2. Timbre, Space and Defaults Attributes

The second part of the test consisted in assessing the quality of contents according to the three attributes: timbre, space and defaults. For each attribute, the original version obtained the higher score (see Figure 4). The method suggested an anchor specific to each attribute, and this anchor was evaluated as the worst item considering the assessment of its associated attribute. For exemple, the timbral anchor "3.5" was the worst item for the timbre attribute, whereas the spatial anchor "SA" was the worst for the spatial attribute. Therefore, assessors seemed to have a good comprehension of attributes. For HEAAC-40 and AMR-48 codings, there is no relevant difference between the attributes. By contrast, the AAC-32 and MP3-64 degradations showed significative differences between the three attributes (see Figure 4). AAC-32 was lower for the timbre and MP3-64 was lower for defaults attribute. Listeners seemed to be able to score in critical maners on each proposed attribute. These observations show the relevance of evaluating specific attributes.

For the spatial attribute evaluation, it was noticed that the spatial anchor "SA" was the lower item but was scored 0.52. It was placed in the middle range of the quality scale. These remark is recurrent in different studies [7] [8]. Figure 5 representes the scores of "SA" focusing on the spatial attribute evaluation for the 7 sequences. Two groups of sequences were clearly highlighted. For three sequences, the spatial anchor was recognized in low quality (<0.4). The second group was hardly affected by the spatial degradation. For spatial evaluation, contents could be an important biais.

4.3. Correlation between overall quality and attributes

A multiple linear regression was carried out in order to quantify the correlation and the weight of sound attributes with the overall quality. The results of correlation analysis are presented in the table 3. The overall quality was correlated to the defaults (0.82), then timbre (0.73) and space (0.52). Correlation between attributes (timbre, space and defaults) were poor.

Results of the regression are summarized in the table 4.

The R value (0.968) and the standard error of the

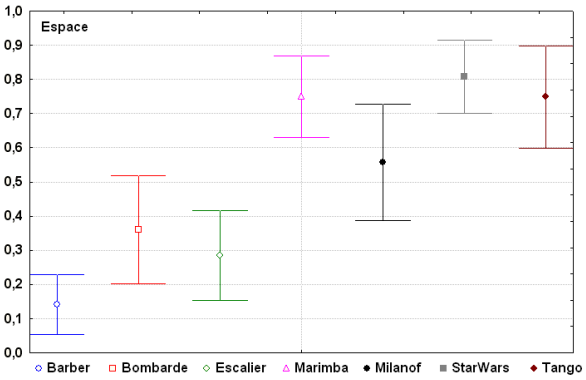


Fig. 5: Spatial anchor scores for spatial attribute evaluation by sequences.

Attribute	Timbre	Space	Defaults
Overall quality	0.73	0.52	0.82
Timbre	-	0.48	0.33
Space	-	-	0.20

Table 3: Correlation values.

estimation indicate that the predicted overall quality and the actual overall quality are very close. The R square value is 0.938 and about 94% of the variance of the overall quality scores can be predicted. Figure 6 representes a scatter plot of the predicted and the observed values of the overall quality. One can note that the regression model denotes a high accuracy.

The aim of this study was to verify the weight of each attribute on the overall quality score. The values of the standardized regression coefficients (β) are 0.41 for timbre, 0.20 for space and 0.64 for defaults attribute which is the most influent attribute on the overall quality (OQ). The coordinates of the regression equation are given by the unstandardized regression coefficient (B) :

$$OQ = 0,67\, defaults + 0,5\, timbre + 0,29\, space - 0.45, \quad (1)$$

These coefficients confirms previous studies which concluded that the timbral fidelity was more influent on the basic audio quality than the spatial fidelity [5]. But the test described in this paper included a

R	R^2	Ajusted R^2	F(3.52)	Std. Error
0.968	0.938	0.934	261.42	0.06

Table 4: Multiple linear regression model summary.

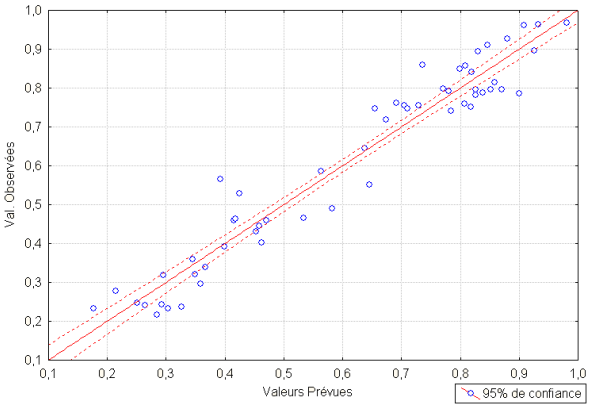


Fig. 6: Observed values vs predicted values for the overall quality.

third attribute named defaults which has a significant impact on the overall quality.

5. DISCUSSION

Choice of anchors.

Test included 3 anchors, each one focused on an attribute, timbre, space and defaults. Results proved that the choice of anchors was appropriate but presented some biais. The spatial anchor was scored in the middle range of the quality scale in different studies [7] [8]. The same observation was noticed in this test. However the analysis according sequences of the spatial anchor focused on the spatial attribute revealed that contents had significant effects in the identification of the spatial anchor. A question appeared about a definition of a spatial anchor so that is scored in low quality for all sequences. Furthermore it is important to keep in mind that this test was run without an explicit reference.

Method adapted for intermediate quality.

Two codings showed relevant differences amoung the three attributes. They were considered as intermediate quality. For the two other codings, considered as higher quality, no difference between the attributes emerged from the analysis. Two explanations can be established, either for high quality it was impossible

to perceive difference between the attributes or the three attributes were equivalently impaired.

6. CONCLUSION

The listening test method proposed in this paper was based on the quality evaluation of three sound attributes, named "timbre, space, defaults". Seven binaural contents were used and 8 degradations were applied (original, 3 anchors, and 4 codings). The method consisted in a multicomparison test but without explicit reference. An anchor specific to each attribute was included. The first session of the test was the overall quality evaluation, then the second stage was the evaluation of the three attributes. The anchors were identified as low quality and thus showed that the attributes were well understood. The spatial anchor was scored as the worst item for spatial attribute but in intermediate quality. Contents can explain this observation, indeed spatial contents strongly affect the spatial impression evaluation. This method dedicated to spatial audio seems to be suitable for intermediate quality assessment and allows to give information about what was impaired. The proposed regression model was accurate and a regression equation was defined. The overall quality could be predicted by scores of the three attributes and "defaults" attribute has more influence than timbre on the overall quality. The space attribute seems to have slight influence on the overall quality assessment.

7. REFERENCES

[1] ITU-R Recommendation BS.1534, "Method for the subjective assessment of intermediate quality level of coding systems," International Telecommunications Union, Radiocommunication Assembly, 2003.

[2] S. Le Bagousse, M. Paquier and C. Colomes, "State of the Art on Subjective Assessment of Spatial Sound Quality," presented at the AES Int. Conf. on Sound Quality Evaluation, Pitea, Sweden, 2010 June –3.

[3] S. Le Bagousse, M. Paquier and C. Colomes, "Families of Sound Attributes for Assessment of Spatial Audio," presented at the AES 129th convention, San Francisco, USA, 2010 October 31 – November 3.

[4] F. Rumsey and S. Zielinski and R. Kassier and S. Bech, "On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio qualitys," presented at the Journal of the Acoustical Society of America, 2005 v.118 n2 pp.968–976.

[5] P. Marin, F. Rumsey and S. Zielinski, "Unraveling the relationship between basic audio quality and fidelity attributes in low bit-rate multichannel audio codecs," presented at the AES 124th convention, Amsterdam, The Netherlands, 2008 May 17–20.

[6] S. Zielinski and P. Brooks and F. Rumsey, "On the use of graphic scales in modern listening tests," presented at the AES 123th convention, New York, USA, 2007 October 5–8.

[7] EBU-TECH 3324, "EBU Evaluations of Multichannel Audio Codecs," European Broascasting Union, 2007.

[8] A. Mason, D. Marston, F. Kozamernik and G. Stoll, "EBU tests of multi-channel audio codecs," presented at the AES 122th convention, Vienna, Austria, 2007 May 5–8.

Spatial audio technologies become very important in audio broadcast services. But, there is a lack of methods for evaluating spatial audio quality. Standards do not take into account spatial dimension of sound and assessments are limited to the overall quality particularly in the context of audio coding. Through different elicitation methods, a long list of attributes has been established to characterize sound but it is difficult to include them in a listening test. A previous study aimed at clustering attributes in families. Thus 3 families of attributes were highlighted, "timbre", "space" and "defects". The overall quality and these three families were evaluated in the listening test presented in this article. The test protocol was based on the Mushra recommendation. However it included three anchors specific to each attribute and no reference in order to evaluate quality instead of fidelity. The aim of the experiment described in this paper was to verify the influence of those 3 attributes on the overall quality in a 5.1 reproduction system. It results that the defects attribute has more influence on the overall quality than the timbre and the space. Moreover the presentation of the three attributes on a same screen adds no bias.

1 Introduction

Before being broadcasted on services, the quality of audio contents has to be evaluated. But, current methods of quality assessment reveal some lacks. Despite the development of spatial technologies, standards do not take into account specific features of spatial sound. The basic audio quality (BAQ) is often the only evaluated attribute. According to ITU-R BS.1534 [1], BAQ is the "global attribute used to judge any and all detected differences between the reference and the object". It would be interesting to obtain some clues on impairments influencing the overall quality. Some attributes, such as coloration, brightness, distortion, localization... have been highlighted by different elicitation methods. However their definitions and their understandings remain a major problem and it is difficult to include them in a listening test [2]. Rather than submitting a list of attributes to the listener, it is possible to gather them in different main sound families. The bias created by specific attributes meanings is therefore reduced. Hence a previous study highlighted 3 sound families for qualifying audio contents: "timbre", "space" and "defects" [3]. Others categories of attributes were defined by studies as timbral, frontal and surround fidelity attributes. These tests showed that timbral fidelity was more correlated to the BAQ than spatial fidelity [4]. For each excerpt, the aim of those experiments was to compare various items to their reference for each of the 4 fidelity parameters. The term fidelity was employed because tests included an explicit reference. One of the requirements for the method tested in this paper was that there were no reference. Nonetheless, the original version, was considered as a hidden reference. The aim of the experiment described in this article was to test a quality evaluation method and to prove the influence, precisely the weight of those attributes families on the overall quality in the context of spatial audio.

2 Attributes families

A previous experiment was run in order to highlight families of sound attributes to evaluate the quality of spatial audio [3]. Tests consisted in presenting a list of attributes (28) and asking assessors to classify them in some categories. No sound was presented in order to create groups independently of audio restitution systems. Two methods were employed: a multidimensional scaling (MDS) and on the other hand a free categorization and a clusters analysis. Both tests obtained the same results and thus three families were defined.

- Defects: are interfering elements or nuisances present in a sound, e.g. noise, distortion, background noise, hum, hiss, disruption

- Space: refers to spatial impression-related characteristics, e.g. depth, width, localization, spatial distribution, reverberation, spatialization, distance, envelopment, immersion

- Timbre: this family is split into 2 subfamilies :
  The first one deals with the sound color, e.g. brightness, tone color, coloration, clarity, hardness, equalization, richness
  The second one composed of homogeneity, stability, sharpness, realism, fidelity and dynamics describes the timbre but can also be related to other characteristics of sound.

3 Listening test

In this study, the 3 attributes, "timbre, space and defects" were included in the listening test.

3.1 Listening conditions

The listening room respected conditions of the recommendation ITU-R BS.1116 [5]. The audio system was a 5.1 restitution system. The five loudspeakers were placed according to the ITU-R BS.775 [6].

3.2 Programme material

Six audio sequences were randomly presented to the assessors. Excerpts were chosen throught film, environment and music to cover a large range of contents. The six sequences were soccer comments, waves and sea sound, movie scene (a fight), music (orchestra, jazz and a turning sound). Each sequence was no longer than twenty seconds according to the recommendation ITU BS.1534 [1]. For each excerpt, six various versions were presented including the original (unprocessed signal), two codecs and three anchors specific to each attributes family. The 6 versions are described in table 1. The spatial anchor was specially defined for this test and was based on anchors used in the litterature [7],[8]. This spatial anchor consisted in a crosstalk between the front right and the surround left channel and the widening of each channel.

"3.5" item was defined as a timbral anchor, "SA" a spatial anchor and "noise" a defects anchor.

Table 1: Description of items.

N°	Abbrevation	Item
1	cod 1	Codec 1
2	3.5	Low pass filtered at 3.5 kHz
3	cod 2	Codec 2
4	noise	Pink noise added
5	o	Original (unprocessed signal)
6	SA	Spatial degradation

3.3 Panel composition

Twenty four "experts" assessors participated in quality tests. They are able to detect impairments in audio signals and they have solid musical background due to their job in audio or musical field. The first test session was made by all the assessor population. However the second part of the test split the panel in two groups.

3.4 Test protocol

The test was decomposed in two sessions. The first one was the evaluation of the overall quality and the second one was the assessment based on the 3 main attributes ("timbre", "space" and "defects"). The test protocol was inspired by Mushra method (ITU-R BS.1534) [1]. Stimuli were presented simultaneously and assessors scored all items on a quality scale. This test included no explicite reference, thought the original version could be considered as an hidden reference. It was noticed that some biases encountered in standards come from the scale [9] thus the proposed grading scale was without labels except on the end point called "low quality" and "high quality". No number appeared during the grading, assessors had to place the cursor along the slider. One instruction was given: the stimulus perceived as the best quality had to be scored at the top of the scale. The interface enabled to zoom on the excerpt for listening smaller part of the whole audio stimulus. First, all the listeners evaluated the overall quality (OQ). Then, eleven of them assessed the 3 attributes (timbre, space and defects) in a same time (see Figure 1) whereas the other group evaluated each attribute one after the other in three succesive subsessions (see Figure 2). The aim was to verify the kind of presentation to employ during a listening test. Is the grading affected by the evaluation of the 3 sound families in a same screen? Are the assessors unabled to focus their attention on different attributes as suggested in other studies? [10]

4 Results

The researched interest is focused on the method to evaluate audio quality rather than the score and the ranking of the sequences and processes.

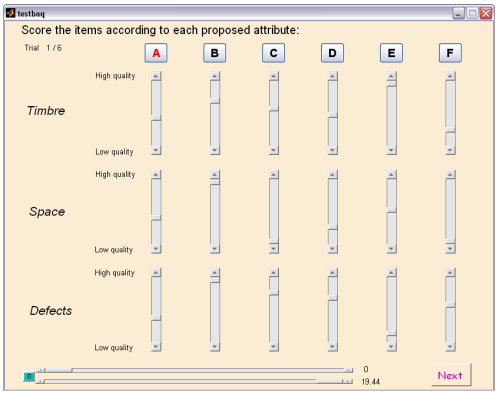


Figure 1: Interface for three attributes presentation

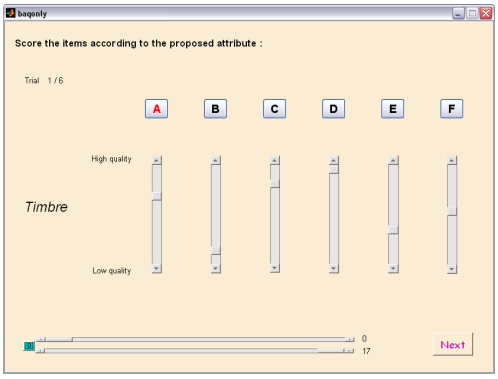


Figure 2: Interface for one attribute presentation

4.1 Attributes presentation

The first thing to notice was the total duration taken by the assessors to complete the test. The single attribute presentation lasted on average 73 minutes whereas the other session took 53 minutes.

Results of the two groups of assessors were compared. A Student test was used to verify the similarity between the scores of both groups. Thus the method of attributes presentation was considered as statistically equivalent. Results obtained by both methods could be merged for the following analyses. Figure 3 shows the similarity between the two methods (method 1 : one attribute, method 2: 3 attributes presentation) for the scoring of timbre attribute (mean scores and error bars show 95% confidence interval).

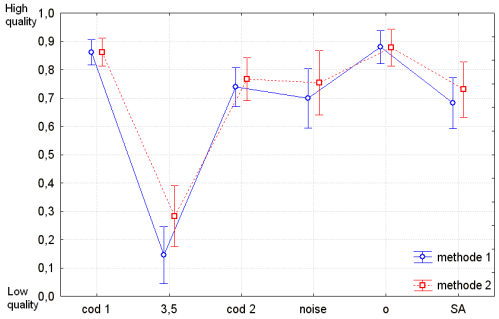


Figure 3: Mean scores and 95% CI of the timbre evaluation for each presentation method

4.2 ANOVA

An ANOVA on each attribute was conducted to highlight factors influencing the scoring. This statistical technique confirmed that the method of attributes presentation had no impact on grading.

Significant effects were revealed by degradations ($p<0.0001$) for each attribute overall quality (OQ), timbre, space and defects). Sequences presented significant effects only for timbre ($F=2.6$, $p=0.032$) and space ($F=7.76$, $p<0.0001$) attributes. Post hoc Tukey's HSD (Honestly Significative Difference) showed that this effect was due to sequences individually and not to a kind of contents (musical and the others excerpts). For exemple, the sequences "sea" and "soccer" were statistically different for spatial attribute whereas they were statistically similar for timbre attribute.

A Tukey's HSD test was performed on degradations for the 4 evaluated attributes. The original version and the "cod 1" had high values and thus were statistically equivalent for all attributes (Tukey values, OQ: 0.96 ; timbre: 0.995 ; space: 0.998 ; defects: 1). By contrast, the rating of the timbral anchor "3.5" is significantly different from the other items for each attribute analysis. For timbre analysis, scores for items "noise", "cod 2" and "SA" were statistically similar. For space analysis, "cod 2" and "noise" ratings are statistically similar with an HSD value of 0.997 and for defects attribute, "cod 2" and "SA" the value was 0.977. Hence, two groups of items were statistically highlighted. The first one consists in the original and the "cod 1" and the second one is composed of "cod 2", "noise" and "SA" but it is dependent on the attributes. An anchor was statistically different from the others items considering the analysis of its associated attribute.

Figure 4 represents mean scores and 95% confidence interval for each attribute evaluation for each item. For both evaluated codings, the obtained notes for each attribute are very close. For exemple, mean values of "cod 2" for all sequences are OQ: 0.72 , timbre: 0.75 , space: 0.76 , defects: 0.76. Moreover "cod 1" is assessed between 0.8 and 0.9 and "cod 2" at about 0.75 for all attributes. The quality of codings used in this test was too high to be included in a test method based on attributes. The overall quality seemed to be sufficient for the assessment of small impairments. With low or intermediate qualities, listeners would be able to detect differences among attributes.

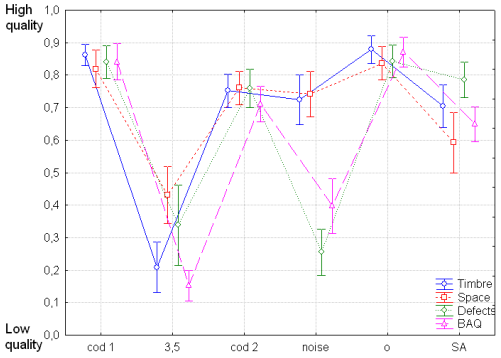


Figure 4: Mean scores and 95% CI of the 4 attributes

4.3 Choice of anchors

Test included 3 anchors, each one focused on an attribute. For the timbre evaluation, the 3.5 item was scored logically in low quality. However this item was also scored in the lower half of the scale for space and defects attributes. The low pass filtered at 3.5 KHz seemed to affect many aspects of sound including space and defects and not only the timbre. The "noise" anchor was the worse item for defects attribute and by contrast, it was scored in high quality for the other attributes (timbre and space). Hence it could be considered as a good anchor for the defects attribute. For the spatial attribute, the spatial anchor (SA) was scored better than timbral anchor ("3.5") and placed in the middle range of the scale, not in low quality (see figure 4). In an other study, spatial anchor was placed in the middle of the quality scale [8]. A question appeared about the possibility to define a spatial anchor scored in low quality. Furthermore it is important to remind that this test was run without explicit reference.

4.4 Correlation between overall quality and attributes

A multiple linear regression was carried out in order to quantify the correlation and the weight of sound attributes with the overall quality.

The results of correlation analysis are presented in the table 2. All variables were correlated with each other. The overall quality was more correlated to the defects (0.90), then timbre (0.87) and space (0.78). Defects attribute was less bonded to the space (0.49) than to the timbre (0.64). Timbre was correlated to space (0.88).

Table 2: Correlation values between overall quality and attributes.

Attributes	Timbre	Space	Defects
Overall quality	0.87	0.78	0.9
Timbre	-	0.88	0.64
Space	-	-	0.49

Results of the regression are summarized in the table 3. The R value (0.985) and the standard error of the estimation indicate that the predicted overall quality and the actual overall quality are very close. The R square value is 0.967 and thus, about 97% of the variance of the overall quality scores can be predicted. Figure 5 representes a scatter plot of the predicted and the observed values of the overall quality. Thus the regression model denotes a high accuracy.

Table 3: Multiple linear regression model summary.

R	R^2	F(3.32)	Std Error of the estimate
0.985	0.97	344.22	0.05

The aim of this study was to find the weight of each attribute on the overall quality score. The values of the standardized regression coefficients (β) are 0,25 for timbre and space attributes and 0.61 for defects which is the most influent attribute on the overall quality (OQ). The coordinates of the regression equation are given by the unstandardized regression coefficients:

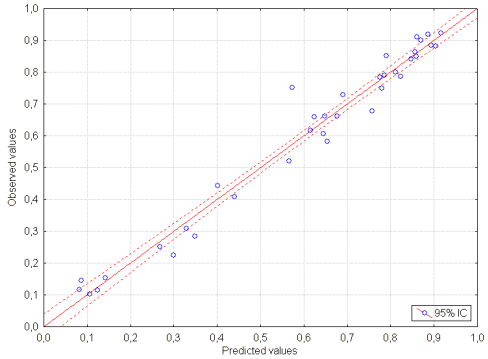


Figure 5: Overall quality, observed vs predicted values

$$OQ = 0,65\ defects + 0,44\ space + 0,3\ timbre - 0.32, \quad (1)$$

These coefficients diverge according to previous studies which concluded that the timbral fidelity was more influent on the basic audio quality than the spatial fidelity [4]. The difference can be explained by the limited number of codings. By consequences, anchors affects strongly the results. Furthermore, the quality was evaluated instead of fidelity (test with no reference). A third attribute called "defects" was introduced and is assessed as the more influent on the overall quality.

5 Conclusion

The listening test method proposed in this paper, was based on the quality evaluation of three sound families, named "timbre, space and defects". Two attributes presentations were tested by assessors, the evaluation of the three attributes simultaneously in one session or the evaluation of attribute in three subsessions sucessively. Results showed that the kind of attributes presentation was not significant. But the three attributes presentation had the advantage of a shorter duration to complete the test. The method included one anchor by attribute. This allowed to verify the well understanding of the attributes definitions by the assessors. The anchors had to be scored in low quality. As mushra test, this method seems to be dedicated to audio with intermediate quality. Impairments on each attribute had to be detected by listeners in order to scores reveal information. The number of evaluated codings was limited. More codecs should be included in the test in order to provide more conclusions. The regression model proposed was accurate. A regression equation was defined and the overall quality could be predicted. This demonstrated the influence of the defects rather than space and timbre on the overall quality. Taking into account those results, a spatial anchor has to be defined and codecs with intermediate quality will be evaluated. Moreover, in the same way, the method is used on headphones with binaural materials.

References

[1] ITU-R Recommendation BS.1534, "Method for the subjective assessment of intermediate quality level of coding systems," International Telecommunications Union, Radio-communication Assembly, Tech. Rep., 2003.

[2] S. Le Bagousse, M. Paquier and C. Colomes, "State of the art on subjective assessment of spatial sound quality," *AES Int. Conf. on Sound Quality Evaluation*, 2010.

[3] S. Le Bagousse, M. Paquier, and C. Colomes, "Families of sound attributes for assessment of spatial audio," *AES 129th Convention*, 2010.

[4] P. Marin, F. Rumsey, and S. Zielinski, "Unraveling the relationship between basic audio quality and fidelity attributes in low bit-rate multi-channel audio codecs," *AES 124th Convention*, 2008.

[5] ITU-R Recommendation BS.1116, "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," International Telecommunications Union, Radio-communication Assembly, Tech. Rep., 1997.

[6] ITU-R Recommendation BS.775, "Multichannel stereophonic sound system with and without accompanying picture," International Telecommunications Union, Radio-communication Assembly, Tech. Rep., 1994.

[7] EBU-TECH 3324, "EBU evaluations of multichannel audio codecs," European Broascasting Union, Tech. Rep., 2007.

[8] A. Mason, D. Marston, F. Kozamernik, and G. Stoll, "EBU tests of multi-channel audio codecs," *AES 122th Convention*, 2007.

[9] S. Zielinski, P. Brooks, and F. Rumsey, "On the use of graphic scales in modern listening tests," *AES 123th Convention*, 2007.

[10] F. Rumsey, S. Zielinski, R. Kassier, and S. Bech, "On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality," *Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 968–976, 2005.

Table des figures

Liste des tableaux

Bibliographie

Algazi, V. R., Avendano, C., et Duda, R. O. (**2001**). "Elevation localization and head-related transfer function analysis at low frequencies", J. Acoust. Soc. Am. **109**, pp. 1110–1122.

Algazi, V. R., Duda, R. O., Duraiswami, R., Gumerov, N. A., et Tang, Z. (**2002**a). "Approximating the head-related transfer function using simple geometric models of the head and torso", J. Acoust. Soc. Am. **112**, pp. 2053–2064.

Algazi, V. R., Duda, R. O., et Thompson, D. M. (**2002**b). "The use of head-and-torso models for improved spatial sound synthesis", AES 113th Convention .

Autti, H. et Biström, J. (**2004**). "Mobile audio–from mp3 to aac and further", Multimedia Seminar : Mobile Multimedia Application Platforms .

Batteau, D. W. (**1967**). "The role of pinna in human localization", Proc. R. Soc. London 158–180.

Bech, S. (**1999**). "Methods for subjective evaluation of spatial characteristics of sound", in *presented at the 16th AES International Conference on Spatial Sound Reproduction* (Rovaniemi, Finland).

Berg, J. (**2005**). "Opaque - a tool for the elicitation and grading of audio quality attributes", presented at the 118th AES Convention .

Berg, J. (**2006**). "Evaluation of perceived spatial audio quality", Journal of Systemics, Cybernetics and Informatics **4**, 10–14.

Berg, J. et Rumsey, F. (**1999**). "Spatial attribute identification and scaling by repertory grid technique and other methods", in *presented at the 16th AES International Conference on Spatial Sound Reproduction* (Rovaniemi, Finland).

Berg, J. et Rumsey, F. (**2000**). "In search of the spatial dimensions of reproduced sound : Verbal protocol analysis and cluster analysis of scaled verbal descriptors", in *presented at the 108th AES Convention* (Paris).

Berg, J. et Rumsey, F. (**2001**). "Verification and correlation of attributes used for describing the spatial quality of reproduced sound", in *presented at the 19th AES International Conference on Surround Sound* (Bavaria, Finland).

Berg, J. et Rumsey, F. (**2003**). "Systematic evaluation of perceived spatial quality", in *presented at the 24th AES International Conference on Multichannel Audio* (Banff).

Borg, I. et Groenen, P. (**2005**). *Modern multidimensional scaling : Theory and applications* (Springer Verlag).

Brandenburg, K. (**1999**). “Mp3 and aac explained”, in *presented at the 17th International Conference on High-Quality Audio Coding* (Florence).

Busson, S. (**2006**). “Individualisation d’indices acoustiques pour la synthèse binaurale”, Ph.D. thesis, Université de la Méditerranée Aix-Marseille II, France.

Carroll, J. et Chang, J. (**1970**). “Analyses of individual differences in multidimensional scaling via an n-way generalization of ’eckart-young’ decomposition”, Psychometroka **35**, 283–319.

Cheminée, P. et Dubois, D. (**2009**). *Le sentir et le dire : concepts et méthodes en psychologie et linguistique cognitives* (L’Harmattan).

Choisel, S. et Wickelmaier, F. (**2006**). “Extraction of auditory features and elicitation of attributes for the assessment of multichannel reproduced sound”, J. Audio Eng. Soc **54**, 815–826.

Choisel, S. et Wickelmaier, F. (**2007**). “Evaluation of multichannel reproduced sound : Sacling auditory attributes underlying listener preferenceextraction of auditory features and elicitation of attributes for the assessment of multichannel reproduced sound”, J. Acoust. Soc. Am. **121**, 388–400.

Cossette, P. (**2004**). *L’organisation : Une perspective cognitiviste* (Les Presses de l’Université Laval).

Dancey, C. P., Reidy, J., et traduction de Gauvrit, N. (**2007**). *Statistiques sans maths pour psychologues : SPSS pour Windows, QCM et exercices corrigés* (De Boeck).

Doignon, J. et Falmagne, J. (**1998**). *Knowledge Spaces* (Springer).

Duda, R. O. et Martens, W. L. (**1998**). “Range dependence of the response of a spherical head model”, J. Acoust. Soc. Am. **104**, pp. 3048–3058.

EBU (**2000**a). “Ebu report on the subjective listening tests of some commercial internet audio codecs. document bpn 029”, Technical Report, Europeen Broadcasting Union.

EBU (**2000**b). “Mushra - method for subjective listening tests of intermediate audio quality. draft ebu recommandation, b/aim 022 (rev.8)/bmc 607rev”, Technical Report, Europeen Broadcasting Union.

EBU Tech 3276 (**1997**). “Listening conditions for the assessment of sound programme material : monophonic and two-channel stereophonic”, Technical Report, Europeen Broadcasting Union.

EBU Tech 3286 (**1997**). “Assessment methods for the subjective evaluation of the quality of sound programme material - music”, Technical Report, Europeen Broadcasting Union.

EBU Tech 3286 Supp. 1 (**2000**). “Assessment methods for the subjective evaluation of the quality of sound programme material - multichannel”, Technical Report, Europeen Broadcasting Union.

Etame, T. (**2008**). “Conception de signaux de référence pour l’évaluation de la qualité perçue des codeurs de la parole et du son”, Ph.D. thesis, Université de Rennes 1.

Gabrielsson, A. et Sjögren, H. (**1979**). "Perceived sound quality of sound reproduction systems", J. Acoust. Soc. Am. **65**, 1019–1033.

Ganter, B., Wille, R., et Franzke, C. (**1997**). *Formal Concept Analysis : Mathematical Foundations* (Springer).

Grey, J. (**1977**). "Multidimensional perceptual scaling of music timbres", J. Acoust. Soc. Am. **61**, 122–135.

Guastavino, C. et Katz, B. (**2004**). "Perceptual evaluation of multi-dimensional spatial audio reproduction", J. Acoust. Soc. Am. **116**, 1105–1115.

Guillon, P. (**2009**). "Individualisation des indices spectraux pour la synthèse binaurale : recherche et exploitation des similarités inter-individuelles pour l'adaptation ou la reconstruction de hrtf", Ph.D. thesis, Université du Maine, Le Mans, France.

Guski, R. (**1997**). "Psychological methods for evaluating sound quality and assessing acoustic information", Acta Acustica **83**, 765–774.

Hebrank, J. et Wright, D. (**1974**). "Spectral cues used in the localization of sound sources on the median plane", J. Acoust. Soc. Am. **56**, pp. 1829–1834.

Heller, J. (**2000**). "Representation and assessment of individual semantic knowledge", Method of Psychological Research **5**, 1–37.

Hofman, P., Riswick, J. V., et Opstal, A. V. (**1998**). "Relearning sound localization with new ears", Nature Neuroscience **1**, pp. 417–421.

ISO.389 (**1985**). "Reference zero for the calibration of pure air tone conduction audiometers", Technical Report, International Organization for Standardization.

ISO/IEC 11172-3 (**1993**). "Information technology - coding of moving pictures and associated audio for digital storage media at up to about 1,5 mbit/s - part 3 : Audio", International Organization for Standardization .

ISO/IEC 13818-7 (**2006**). "Information technology - generic coding of moving pictures and associated audio - part 7 : Advanced audio coding (aac)", International Organization for Standardization .

ISO/IEC 14496-3 (**2005**). "Information technology - coding of audio-visual objects - part 3 : Audio", International Organization for Standardization .

ITU-R BS.1116 (**1997**). "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems", Technical Report, International Telecommunications Union, Radio-communication Assembly.

ITU-R BS.1284 (**2003**). "Method for the subjective assessment of intermediate quality level of coding systems", Technical Report, International Telecommunications Union, Radio-communication Assembly.

ITU-R BS.1534 (**2003**). "Method for the subjective assessment of intermediate quality level of coding systems", Technical Report, International Telecommunications Union, Radio-communication Assembly.

ITU-R BS.775-2 (**2006**). "Multichannel stereophonic sound system with and without accompanying picture", Technical Report, International Telecommunications Union, Radio-communication Assembly.

ITU-R BT.1082-1 (**1990**). "Studies toward the unification of picture assessment methodology", Technical Report, International Telecommunications Union, Radio-communication Assembly.

ITU-R BT.500-11 (**2002**). "Methodology for the subjective assessment of the quality of television picture", Technical Report, International Telecommunications Union, Radio-communication Assembly.

ITU-T G.722-2 (**2002**). "Wideband coding of speech at around 16 kbit/s using adaptive multi-rate wideband (amr-wb)", Technical Report, International Telecommunications Union, Radio-communication Assembly.

ITU-T P.910 (**1999**). "Subjective video quality assessment methods for multimedia applicationsmethod for the subjective assessment of intermediate quality level of coding systems", Technical Report, International Telecommunications Union, Radio-communication Assembly.

Jones, B. et McManus, P. (**1986**). "Graphic scaling of qualitative terms", SMPTE Journal **95**, 1166–1171.

Kelly, G. (**1955**). *The psychology of personal constructs* (Norton, New York).

Kistler, D. J. et Wightman, F. L. (**1992**). "A model of head related transfer function based on principal components analysis and minimum-phase reconstruction", J. Acoust. Soc. Am. **91**, pp. 1637–1647.

Koehl, V. et Paquier, M. (**2013**). "A comparative study on different assessment procedures applied to loudspeaker sound quality", Applied Acoustics **74**, pp. 1448–1457.

Koivuniemi, K. et Zacharov, N. (**2001**). "Unraveling the perception of spatial sound reproduction : Language development, verbal protocol analysis and listener training", in *presented at the 111th AES Convention* (New York).

Kruskal, J. B. (**1964**). "Nonmetric multidimensional scaling : a numerical method", Psychometrika **29**, 115–129.

Kruskal, J. B. et Wish, M. (**1978**). *Multidimensional Scaling* (Sage Publications).

Kulkarni, A., Isabelle, S. K., et Colburn, H. (**1995**). "On the minimum-phase approximation of head-related transfer functions", IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) .

Lavandier, M., Herzog, P., et Meunier, S. (**2008**a). "Comparative measurements of loudspeakers in a listening situation", J. Acoust. Soc. Am. **123**, 77–87.

Lavandier, M., Meunier, S., et Herzog, P. (**2008**b). "Identification of some perceptual dimensions underlying loudspeaker dissimilarities", J. Acoust. Soc. Am. **123**, 4186–4198.

Lavandier, M., Meunier, S., Herzog, P., *et al.* (**2005**). "Perceptual and physical evaluation of differences among a large panel of loudspeakers", in *Forum Acusticum.*

Lawless, H. et Heymann, H. (**1998**). *Sensory evaluation of food : priciples and practices* (Chapman and Hall).

Le Bagousse, S., Colomes, C., et Paquier, M. (**2010**). "State of the art on subjective assessment of spatial audio quality", in *presented at the 38th AES International Conference on Sound Quality Evaluation* (Pitea).

Le Bagousse, S., Paquier, M., et Colomes, C. (**2012**). "Assessment of spatial audio quality based on sound attributes", in *presented at Acoustics 2012* (Nantes).

Letowski, T. (**1989**). "Sound quality assessment : Concepts and criteria", presented at the 87th AES Convention .

Lorho, G. (**2005**a). "Evaluation of spatial enhancement systems for stereo headphone reproduction by preference and attribute rating", in *presented at the 118th AES Convention* (Barcelona).

Lorho, G. (**2005**b). "Individual vocabulary profiling of spatial enhancement systems for stereo headphone reproduction", in *presented at the 119th AES Convention* (New York).

Lorho, G. (**2010**). "Perceived quality evaluation an application to sound reproduction over headphones", Ph.D. thesis, School of science and technology, Aalto Univeristy.

Makinen, J., Bessette, B., Bruhn, S., Ojala, P., Salami, R., et Taleb, A. (**2005**). "Amr-wb+ : a new audio coding standard for 3rd generation mobile audio services", in *IEEE ICASSP*, volume 2, 1109–1112 (Philadelphia).

Marins, P., Rumsey, F., et Zielinski, S. (**2006**). "The relationship between selected artifacts and basic audio quality in perceptual audio codecs", in *presented at the 120th AES Convention* (Paris).

Marins, P., Rumsey, F., et Zielinski, S. (**2007**). "The relationship between basic audio quality and selected artefacts in perceptual audio codecs - part ii : Validation experiment", in *presented at the 122th AES Convention* (Vienna).

Marins, P., Rumsey, F., et Zielinski, S. (**2008**). "Unravelling the relationship between basic audio quality and fidelity attributes in low bit-rate multi-channel audio codecs", in *presented at the 124th AES Convention* (Vienna).

Mason, A., Marston, D., Kozamernik, F., et Stoll, G. (**2007**). "Ebu tests of multi-channel audio codecs", in *presented at the 122th AES Convention* (vienna).

Meltzer, S. et Moser, G. (**2006**). "He-aac v2 mpeg-4 - audio coding for today's digital media world", Technical Report, EBU Technical Review.

Moulin, S. (**2011**). "Évaluation subjective de la qualité sonore en écoute binaurale", Master's thesis, Université du Maine.

Naes, T. et Risvik, E. (**1996**). *Multivariate analysis of data in sensory science* (Elsevier Science Ldt).

Nakayama, T., Miura, T., Kosaka, O., Okamoto, M., et Shiga, T. (**1971**). "Subjective assessment of multichannel reproduction", J. Audio Eng. Soc. **19**, 744–751.

Narita, N. (**1993**). "Graphic scaling and validity of japanese descriptive terms used in subjective evaluation tests", SMPTE Journal **102**, 616–622.

Pedersen, T. et Zacharov, N. (**2008**). "How many psycho-acoustic attributes are needed ?", in *presented at Acoustics 08* (Paris).

Pernaud, J. M. (**2003**). "Spatialisation du son par les techniques binaurales : Application aux services de télécommunications", Ph.D. thesis, Institut National Polytechnique de Grenoble, France.

Pernaux, J. (**2003**). "Spatialisation du son par les techniques binaurales : application aux services de télécommunications", Ph.D. thesis, Institut national polytechnique de Grenoble.

Rayleigh, L. (**1907**). "On our perception of sound direction", Philisophical Magazine **13**, pp. 214–232.

Rumsey, F. (**1998**). "Subjective assessment of the spatial attributes of reproduced sound", in *presented at the 15th AES International Conference on Audio, Acoustics & Small Spaces* (Copenhagen, Denmark).

Rumsey, F., Zielinski, S., Kassier, R., et Bech, S. (**2005**). "On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality", J. Acoust. Soc. Am. **118**, 968–976.

Schatz, R., Egger, S., et Masuch, K. (**2012**). "The impact of test duration on user fatigue and reliability of subjective quality ratings", J. Audio Eng. Soc **60**, 63–73.

Shaw, E. et Teranishi, R. (**1968**). "Sound pressure generated in an external-ear replica and real human ears by a nearby point source", J. Acoust. Soc. Am. **44**, pp. 240–249.

Soulodre, G. et Lavoie, M. (**1999**). "Subjective evaluation of large and small impairments in audio codecs", in *presented at the 17th AES International Conference on High-Quality Audio Coding.*

Susini, P., McAdams, S., et Winsberg, S. (**1999**). "A multidimensional technique for sound quality assessment", Acta Acustica **85**, 650–656.

Takane, Y., Young, F., et Leeuw, J. D. (**1977**). "Nonmetric individual differences multidimensional scaling : an alternating least squares method with optimal scaling features", Psychometrika **42**, 7–67.

Teunissen, K. (**1996**). "The validity of ccir quality indicators along a graphical scale", SMPTE journal **105**, 144–149.

Toole, F. (**1985**). "Subjective measurements of loudspeaker sound quality and listener performance", J. Audio Eng. Soc. **33**, 2–32.

Tournois, J. et Dickes, P. (**1993**). *Pratique de l'échelonnement multidimensionnel : de l'observation à l'interprétation* (De Boeck).

Virtanen, M., Gleiss, N., et Goldstein, M. (**1995**). "On the use of evaluative category scales in telecommunications", in *Proceedings of Human Factors in Telecommunications* (Melbourne).

Ward, J. (**1963**). “Hierarchical grouping to optimize an objective function”, J. Am. Stat. Assoc. **58**, 236–244.

Watson, A. (**1999**). “Assessing the quality of audio and video components in desktop multimedia conferencing”, Ph.D. thesis, Universite college London.

wiktionary (**2014**). URL `http://fr.wiktionary.org/wiki/stéréo-`.

Wolters, M., Kjorling, K., Homm, D., et Purnhagen, H. (**2003**). “A closer look into mpeg-4 high efficiency aac”, in *presented at the 115th AES Convention* (New York).

Yannou, B. et Deshayes, P. (**2006**). *Intelligence et innovation en conception de produits et services* (Editions L’Harmattan).

Zielinski, S. (**2006**). “On some biases encountered in modern listening tests”, in *Spatial audio & sensory evaluation techniques conference* (GuilfordRovaniemi, UK).

Zielinski, S., Brooks, P., et Rumsey, F. (**2007**a). “On the use of graphic scales in modern listening tests”, in *presented at the 123th AES Convention* (New York).

Zielinski, S., Hardisty, P., Hummersone, C., et Rumsey, F. (**2007**b). “Potential biases in mushra listening tests”, in *presented at the 123th AES Convention* (New York).

Zielinski, S., Rumsey, F., et Bech, S. (**2003**). “Comparison of quality degradation effects caused by limitation of bandwidth and by down-mix algorithms in consumer multichannel audio delivery systems”, in *presented at the 114th AES Convention* (Amsterdam).

Zielinski, S., Rumsey, F., et Bech, S. (**2008**). “On some biases encountered in modern audio quality listening tests - a review”, J. Audio Eng. Soc **56**, 427–451.

Zielinski, S. K., Rumsey, F., Kassier, R., et Bech, S. (**2005**). “Comparison of basic audio quality and timbral and spatial fidelity changes caused by limitation of bandwidth and by down-mix algorithms in 5.1 surround audio systems”, J. Audio Eng. Soc **53**, 174–192.