Quel son spatialisé pour la vidéo 3D ? : influence d'un rendu Wave Field Synthesis sur l'expérience audio-visuelle 3D

Samuel Moulin

Thèse de doctorat, Université Paris Descartes

Spécialité : Sciences cognitives

Quel son spatialisé pour la vidéo 3D ?

Influence d'un rendu Wave Field Synthesis sur l'expérience audio-visuelle 3D.

Samuel Moulin

Ecole Normale Supérieure - Laboratoire des Systèmes Perceptifs
Ecole Doctorale 261 : Cognition, Comportement, Conduites Humaines
Université Paris Descartes

Thèse réalisée à Orange Labs, Lannion
Version du 11/02/2015

Jury :

Hervé LISSEK	EPFL, Lausanne	Rapporteur
Isabelle VIAUD-DELMON	IRCAM, Paris	Rapporteur
Gilles COPPIN	Télécom Bretagne, Brest	Examinateur
Patrick LE CALLET	Ecole Polytechnique, Nantes	Examinateur
Clara SUIED	IRBA, Brétigny-sur-Orge	Examinatrice
Laetitia GROS	Orange Labs, Lannion	Encadrante
Rozenn NICOL	Orange Labs, Lannion	Encadrante
Pascal MAMASSIAN	Ecole Normale Supérieure, Paris	Directeur de thèse

Laboratoire des Systèmes Perceptifs

À toi, Pierre.

Remerciements

Je tiens, tout d'abord, à témoigner ma gratitude envers Hervé Lissek et Isabelle Viaud-Delmon pour avoir accepté d'être rapporteurs de ce travail de thèse, sans oublier Gilles Coppin, Patrick Le Callet et Clara Suied, pour leur rôle d'examinateurs.

Je remercie également Orange Labs et Bruno Lozach de m'avoir accueilli au sein de l'équipe TPS. Je suis très reconnaissant des moyens mis à ma disposition, grâce auxquels j'ai eu la chance de mettre en œuvre mes expériences dans d'excellentes conditions. Le soutien d'Orange et de l'Université Paris Descartes ont également rendu possible la diffusion de mes travaux, lors de congrès notamment.

Je tiens tout particulièrement à remercier Pascal Mamassian d'avoir dirigé mes travaux de recherche. Nos rencontres ont toujours été enrichissantes et très chaleureuses. Merci, bien évidemment, à Laetitia Gros et Rozenn Nicol pour leur travail d'encadrement au sein d'Orange Labs, mais aussi et surtout, pour le soutien qu'elles m'ont témoigné tout au long de cette aventure. La complicité, la douceur et les rires ont coloré tous nos échanges. Travailler avec vous fut un réel plaisir.

J'adresse également mes remerciements aux nombreuses personnes qui ont activement participé à ce travail : Thibaut Carpentier et Olivier Warusfel pour le partage de leurs outils de spatialisation WFS, Grégory Pallone pour les mesures acoustiques et nos divers échanges scientifiques, Mickaël Bonin pour son aide lors du développement des interfaces de test, ainsi que Jérôme Fournier, Darya Khaustova et Marina Zannoli pour leurs connaissances sur les technologies vidéo 3D et la perception visuelle. Serge Le Boucher (a.k.a "M. Bricolage") m'a également éclairé dans de nombreuses situations grâce à son expertise informatique et son esprit pratique. Je n'oublie pas Jean-Charles Gicquel, Loic Gourmelon, Jean-Yves Leseure, Julien Libouban, Thierry Le Vacon et Rémi Rouaud pour leur aide précieuse lors du tournage et leur réactivité face à la mise en place rapide de ce projet.

À mes collègues et amis, Julian, Joachim, Pyo, Julien, Jérôme, Stéphane, Felipe, Arnaud, et, bien sûr, aux "filles du $2^{ème}$" Julie et Maty, merci d'avoir rendu les pauses, les repas, les soirées et les répètes toujours plus conviviales.

Un grand merci à mes parents et mes grandes sœurs pour leur soutien de tous les jours et leurs encouragements qui m'ont aidé à avancer tout au long de mes études.

Enfin, merci à toi, Chloé, qui m'as accompagné, apaisé, reboosté tout au long de ces années. Merci aussi pour ta participation active dans ce travail, en endossant successivement le rôle de testeuse, d'actrice, de traductrice. Tu as également relu attentivement ce manuscrit et rectifié mes maladresses de "scientifique". Merci d'être une super coéquipière.

Résumé

Le monde du divertissement numérique connaît depuis plusieurs années une évolution majeure avec la démocratisation des technologies vidéo 3D. Il est désormais commun de visualiser des vidéos stéréoscopiques sur différents supports : au cinéma, à la télévision, dans les jeux vidéo, etc. L'image 3D a considérablement évolué, mais qu'en est-il des technologies de restitution sonore associées ? La plupart du temps, le son qui accompagne la vidéo 3D est basé sur des effets de latéralisation, plus au moins étendus (stéréophonie, systèmes 5.1). Il est pourtant naturel de s'interroger sur le besoin d'introduire des évènements sonores en lien avec l'ajout de cette nouvelle dimension visuelle : la profondeur. Plusieurs technologies semblent pouvoir offrir une description sonore 3D de l'espace (technologies binaurales, Ambisonics, Wave Field Synthesis). Le recours à ces technologies pourrait améliorer la qualité d'expérience de l'utilisateur, non seulement en termes de réalisme, grâce à l'amélioration de la cohérence spatiale audio-visuelle, mais aussi en termes de sensation d'immersion.

Afin de vérifier cette hypothèse, nous avons mis en place un système de restitution audio-visuelle 3D proposant une présentation visuelle stéréoscopique, associée à un rendu sonore spatialisé par Wave Field Synthesis. Trois axes de recherche ont alors été étudiés :

1. **Perception de la distance en présentation unimodale ou bimodale**
   Dans quelle mesure le système audio-visuel est-il capable de restituer des informations spatiales relatives à la distance, dans le cas d'objets sonores, visuels, ou audio-visuels ? Les expériences menées montrent que la Wave Field Synthesis permet de restituer la distance de sources sonores virtuelles. D'autre part, les objets visuels et audio-visuels sont localisés avec plus de précision que les objets uniquement sonores.
2. **Intégration multimodale suivant la distance**
   Comment garantir une perception spatiale audio-visuelle cohérente de stimuli simples ? Nous avons mesuré l'évolution de la fenêtre d'intégration spatiale audio-visuelle suivant la distance, c'est-à-dire les positions des stimuli audio et visuels pour lesquelles la fusion des percepts a lieu.
3. **Qualité d'expérience audio-visuelle 3D**
   Quel est l'apport du rendu de la profondeur sonore sur la qualité d'expérience audio-visuelle 3D ? Nous avons, tout d'abord, évalué la qualité d'expérience actuelle, lorsque la présentation de contenus vidéo 3D est associée à une bande son 5.1, diffusée par des systèmes grand public (système 5.1, casque, et barre de son). Nous avons ensuite étudié l'apport du rendu de la profondeur sonore grâce au système audio-visuel proposé (vidéo 3D associée à la Wave Field Synthesis).

Abstract

The digital entertainment industry is undergoing a major evolution due to the recent spread of stereoscopic-3D videos. It is now possible to experience 3D by watching movies, playing video games, and so on. In this context, video catches most of the attention but what about the accompanying audio rendering? Today, the most often used sound reproduction technologies are based on lateralization effects (stereophony, 5.1 surround systems). Nevertheless, it is quite natural to wonder about the need of introducing a new audio technology adapted to this new visual dimension: the depth. Many alternative technologies seem to be able to render 3D sound environments (binaural technologies, ambisonics, Wave Field Synthesis). Using these technologies could improve users' quality of experience. It could impact the feeling of realism by adding audio-visual spatial congruence, but also the immersion sensation.

In order to validate this hypothesis, a 3D audio-visual rendering system is set-up. The visual rendering provides stereoscopic-3D images and is coupled with a Wave Field Synthesis sound rendering. Three research axes are then studied:

1. **Depth perception using unimodal or bimodal presentations**
   How the audio-visual system is able to render the depth of visual, sound, and audio-visual objects? The conducted experiments show that Wave Field Synthesis can render virtual sound sources perceived at different distances. Moreover, visual and audio-visual objects can be localized with a higher accuracy in comparison to sound objects.
2. **Crossmodal integration in the depth dimension**
   How to guarantee the perception of congruence when audio-visual stimuli are spatially misaligned? The extent of the integration window was studied at different visual object distances. In other words, according to the visual stimulus position, we studied where sound objects should be placed to provide the perception of a single unified audio-visual stimulus.
3. **3D audio-visual quality of experience**
   What is the contribution of sound depth rendering on the 3D audio-visual quality of experience? We first assessed today's quality of experience using sound systems dedicated to the playback of 5.1 soundtracks (5.1 surround system, headphones, soundbar) in combination with 3D videos. Then, we studied the impact of sound depth rendering using the set-up audio-visual system (3D videos and Wave Field Synthesis).

Table des matières

Introduction générale

Contexte et objectifs de ce travail

Nous constatons, depuis de nombreuses années, un fort développement des applications multimédias, avec des services toujours plus immersifs et réalistes proposés aux utilisateurs. Ces évolutions concernent notamment les technologies de l'image à travers l'arrivée de la télévision en couleurs dans les années 60, ou plus récemment de la vidéo 3D. Ce développement porte également sur l'augmentation de la résolution des images (passage à la Haute Définition, au 4K et bientôt à la Ultra Haute Définition). Les technologies de restitution sonore ont également évolué au fil des années dans le but d'offrir une expérience sonore toujours plus immersive, en passant progressivement d'une restitution monophonique à stéréophonique, puis multicanale, en particulier grâce à la démocratisation des systèmes 5.1. Durant la dernière décennie, les systèmes audio 3D se sont considérablement développés avec l'arrivée de systèmes multicanaux plus évolués (système 22.2 par exemple), ou encore de la technologie binaurale. Ces différentes technologies de restitution sonore ou visuelle se retrouvent aujourd'hui dans des applications variées telles que la réalité virtuelle, la réalité augmentée, mais aussi dans certaines applications destinées au grand public comme les jeux vidéo, le cinéma ou la télévision, par exemple.

Suivant les applications considérées, les informations sonores et visuelles sont ainsi diffusées au moyen de systèmes variés, et suivant une ou plusieurs dimensions : dans le plan horizontal, en élévation, mais également en profondeur. Cependant, la sensibilité du système perceptif humain peut varier en fonction de la position des percepts, et surtout en fonction de la modalité stimulée. De plus, la présentation simultanée d'informations audio-visuelles donne lieu à des phénomènes d'interaction bimodale complexes. Il semble donc pertinent, pour le développement de toute application audio-visuelle, de considérer conjointement les technologies de restitution sonore et visuelle mises en présence, afin d'offrir aux spectateurs un certain niveau de cohérence entre les espaces reproduits.

Dans le contexte de ce travail, nous nous intéressons plus particulièrement aux applications utilisant les technologies vidéo 3D. Il est évident que l'apport du relief visuel impacte de manière importante l'expérience audio-visuelle des spectateurs. Cependant, dans les applications destinées au grand public, les technologies audio qui accompagnent ces images 3D sont généralement identiques à celles utilisées pour les vidéos 2D. Il est pourtant naturel de s'interroger sur le besoin d'accompagner cette nouvelle dimension visuelle par une technologie de restitution sonore spécifique. La question sous-jacente est de savoir s'il est possible d'améliorer la qualité d'expérience audio-visuelle 3D grâce à un système sonore offrant une sensation de relief sonore,

analogue au relief visuel. En effet, là où la stéréoscopie apporte des effets de profondeur visuelle, les technologies audio offrent généralement des effets de latéralisation, plus ou moins étendus. Il est donc probable qu'un système audio permettant la restitution d'éléments sonores dans la profondeur puisse rendre l'expérience audio-visuelle plus cohérente. Cela soulève différentes questions :

- Comment l'expérience audio-visuelle 3D proposée grâce aux technologies disponibles actuellement est-elle perçue ? Est-il nécessaire d'introduire une nouvelle technologie de restitution sonore pour accompagner les vidéos 3D ?

- Est-il possible de proposer aux spectateurs une restitution sonore offrant une sensation de distance ? Si oui, grâce à quelle(s) technologie(s) ?

- Quelle est la sensibilité des spectateurs à la restitution de la distance d'objets sonores par rapport au relief d'objets visuels ?

- Dans quelles limites des objets virtuels audio-visuels sont-ils perçus comme étant spatialement cohérents ?

- Est-ce que la restitution d'objets audio-visuels spatialement cohérents suivant la distance peut améliorer la qualité d'expérience des utilisateurs ?

Avant de pouvoir répondre à ces questions, il est nécessaire de comprendre comment notre système perceptif capte et interprète les informations sonores et visuelles de notre environnement. Nous devons également nous intéresser aux différentes technologies audio et vidéo 3D qui sont actuellement utilisées dans les applications audio-visuelles destinées au grand public.

Organisation du manuscrit

Le manuscrit présenté ici est structuré autour de sept chapitres dont le but est de répondre aux différentes questions mentionnées précédemment. Ces travaux portent sur trois axes de recherche principaux : l'évaluation de la qualité d'expérience audio-visuelle 3D (chapitres 3 et 7), la perception de la distance d'objets virtuels (chapitre 5), et les phénomènes d'interaction multimodale qui garantissent la perception de cohérence audio-visuelle (chapitre 6). Les chapitres 1 et 2 ont, quant à eux, pour objectif de présenter les aspects perceptifs et technologiques propres à notre problématique. Le contenu de ces sept chapitres est détaillé ci-après.

— Le **premier chapitre** de ce travail expose les principaux mécanismes perceptifs qui nous permettent, au quotidien, de localiser un objet dans l'espace. Nous nous intéressons dans un premier temps au cas d'un objet unimodal, c'est-à-dire au cas d'une stimulation uniquement sonore ou visuelle, avant d'aborder les phénomènes mis en jeu lors d'une présentation bimodale (stimulation audio-visuelle). Ce chapitre met également en évidence la tolérance (voire la plasticité) de notre système perceptif à travers

divers exemples tels que la fusion audio-visuelle de percepts unimodaux présentant des écarts spatiaux ou temporels, par exemple.

— Le **deuxième chapitre** est, quant à lui, dédié aux aspects technologiques, avec la description des principaux systèmes de restitution sonore capables d'offrir des effets de spatialisation sonore. Le principe de fonctionnement des technologies de rendu stéréoscopique 3D est également présenté. Nous verrons notamment que la restitution du relief dépend de nombreux facteurs comme les paramètres de captation des contenus mais aussi des conditions de visualisation. Enfin, nous verrons comment ces différentes technologies sont actuellement intégrées dans de nombreuses applications audio-visuelles destinées au grand public.

— Le **troisième chapitre** introduit la notion de qualité d'expérience audio-visuelle, qui est généralement étudiée à travers la qualité perçue des contenus audio et vidéo. Ce chapitre présente également la première expérience subjective réalisée dans le cadre de ce travail de thèse. Cette expérience a pour but de dresser un état des lieux de la qualité d'expérience audio-visuelle 3D proposée par les technologies de reproduction audio et vidéo 3D utilisées aujourd'hui. Cette étude de la situation actuelle fait ressortir le lien existant entre la notion d'immersion audio-visuelle et la sensation de cohérence entre l'image et le son. Nous proposons, à partir de cette observation, d'explorer l'apport potentiel de la restitution de scènes sonores spatialement cohérentes avec les éléments visuels.

— Le **quatrième chapitre** est consacré à la description du système de restitution audio-visuelle mis en œuvre dans le cadre de ce travail, qui associe une présentation visuelle stéréoscopique à un rendu sonore spatialisé par Wave Field Synthesis. Les performances acoustiques de ce système, appelé AV3D dans la suite de ce document, sont ensuite évaluées de manière objective.

— Dans le **cinquième chapitre**, nous nous intéressons à la capacité du système AV3D à reproduire la distance d'objets virtuels. Une première expérience est menée sur la localisation en distance de haut-parleurs réels. Cette expérience nous permet de valider le protocole expérimental permettant le recueil des jugements de distance. Trois expériences subjectives sont ensuite réalisées suivant ce protocole expérimental. Les deux premières portent sur la perception de la distance d'objets sonores ou visuels. Les stimuli sont ici présentés en condition unimodale. Ensuite, nous étudions la perception de la distance dans le cas d'objets audio-visuels, c'est-à-dire présentés en condition bimodale.

— Le **sixième chapitre** porte, quant à lui, sur la perception de la cohérence entre des stimuli audio et visuels simulés suivant la distance. Nous souhaitons savoir si la position d'un objet sonore placé devant ou derrière un objet visuel peut biaiser la localisation de ce dernier. Ce phénomène porte le nom d'effet ventriloque inverse. Nous désirons

également connaître les conditions pour lesquelles des objets sonores et visuels placés à des positions différentes sont intégrés par le système perceptif, et interprétés comme un objet audio-visuel unique. Ces conditions garantissent en effet la perception de cohérence audio-visuelle qui semble cruciale pour l'amélioration potentielle de la qualité d'expérience.

— Enfin, dans le **septième chapitre**, l'apport potentiel du rendu de la distance d'objets sonores sur la qualité d'expérience audio-visuelle est étudié. Pour cela, un tournage est dans un premier temps effectué avec l'aide d'une équipe professionnelle. Cette étape permet de générer des séquences audio-visuelles représentatives de contenus réalistes dont les caractéristiques sont maîtrisées (position des personnages, rendu de la profondeur visuelle, pistes audio, etc.). Ces séquences sont ensuite évaluées par des spectateurs lors d'un test subjectif. Différents mixages sonores sont alors présentés pour accompagner les vidéos 3D. Seule la distance des éléments sonores varie suivant les mixages ce qui permet d'estimer l'apport du rendu de la distance des objets sonores sur la qualité d'expérience audio-visuelle 3D.

Chapitre 1

Perception spatiale unimodale et bimodale

1.1 Introduction

L'audition et la vision nous fournissent au quotidien des informations sur l'environnement qui nous entoure. L'objectif de ce chapitre est de présenter les principaux mécanismes perceptifs qui permettent à notre cerveau d'interpréter ces informations. Dans cette section, nous allons principalement nous intéresser à la perception spatiale.

Après une description des systèmes perceptifs auditif et visuel (capteurs d'informations unimodales), nous allons aborder la notion de perception bimodale. En effet, la plupart du temps, les informations sonores et visuelles interagissent et sont interprétées comme un percept unique par notre cerveau.

1.2 Perception spatiale sonore

Dans cette section, nous allons considérer le repère sphérique (r, θ, ϕ) décrivant l'espace à trois dimensions (figure 1.1). Les indices visuels et auditifs permettant de localiser les sources dans cet espace tridimensionnel (distance r, azimut θ, élévation ϕ) sont décrits ci-après.

1.2.1 Perception de la direction d'une source sonore

Les indices interauraux

La théorie Duplex proposée par Rayleigh [Rayleigh, 1907] permet de décrire de manière simple les mécanismes de perception auditive lorsque la source sonore est située dans le plan horizontal (localisation en azimut). Cette théorie est basée sur le constat suivant : grâce à leur position différente, nos deux oreilles captent des signaux sonores distincts représentatifs d'un évènement sonore unique. Ces différences permettent de définir deux indices interauraux : la différence interaurale de temps ou ITD (*Interaural Time Difference*), et la différence interaurale d'amplitude ou 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. Pour une source située en dehors du plan médian, le

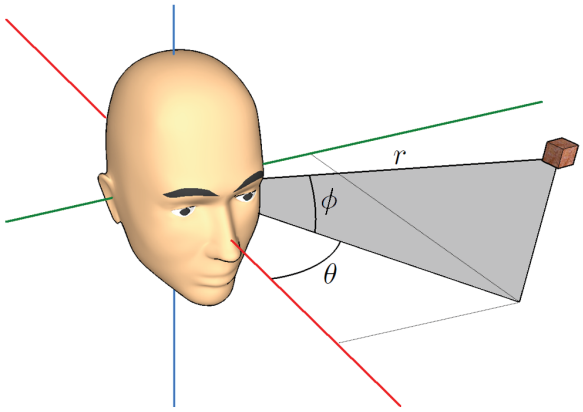


FIGURE 1.1 – *Repère sphérique où r représente la distance, θ l'azimut, et ϕ l'élévation de la source.*

retard interaural occasionné est la conséquence de la différence entre le trajet "source – oreille ipsilatérale" et le trajet "source - oreille controlatérale". L'ITD dépend de la longueur d'onde, c'est-à-dire de la fréquence de l'onde acoustique incidente. L'oreille humaine est sensible à la différence de phase pour les signaux sinusoïdaux ou périodiques de fréquences inférieures à 800 Hz environ [Zwislocki and Feldman, 1956], et au retard d'enveloppe pour les fréquences supérieures. Ainsi, l'ITD porte l'information du degré de latéralisation de la source sonore et ce jusqu'à une fréquence d'environ 1500 Hz [Blauert, 1997, Moore, 2003]. Il faut noter que cette fréquence limite est dépendante de la distance interaurale, qui est d'environ 17 cm chez l'adulte. Le modèle de Woodworth [Woodworth and Schloesberg, 1962] permet d'estimer de manière simplifiée les variations de l'ITD en fonction de l'azimut θ grâce à l'équation

$$ITD_{HF}(\theta) = \frac{a}{c}(sin(\theta) + \theta), \tag{1.1}$$

où a est le rayon d'une tête approximée à une sphère, et c la célérité des ondes acoustiques.

Le second indice perceptif interaural est l'ILD (Interaural Level Difference). Il caractérise les différences de niveau sonore entre les deux oreilles pour une source à une position donnée. Ces différences sont induites par la présence de la tête dans le champ sonore. Cette dernière peut alors être vue comme un obstacle qui occasionne des phénomènes de diffraction plus ou moins importants en fonction de la direction et de la fréquence de l'onde incidente. Aux basses fréquences, ces effets sont faibles car la longueur d'onde acoustique est grande devant la dimension interaurale. Lorsque la longueur d'onde diminue, et donc que la fréquence augmente, les effets de réflexion/diffraction par la tête sont de plus en plus prononcés. L'ILD permet ainsi une perception du degré de latéralisation pour les fréquences supérieures à environ 1500 Hz.

Les indices spectraux

La théorie Duplex proposée par 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. Par exemple, les différences interaurales sont quasi-inexistantes dans le cas de sources situées dans le plan médian. Cependant, celles-ci sont localisables grâce à des indices auditifs supplémentaires : les indices spectraux (IS).

Ces indices prennent en compte les effets des réflexions et diffractions causés par l'interaction des ondes sonores avec le corps de l'auditeur (pavillons, torse, épaules, etc.). Ils peuvent être vus comme une modification spectrale du signal source par la morphologie du sujet. Il est communément admis que l'auditeur interprète ce filtrage anatomique à chaque situation donnée et le compare avec les filtrages stockés en mémoire [Middlebrooks, 1992, Guillon, 2009]. Il peut ensuite en déduire la position de la source par reconnaissance de forme. Les indices spectraux permettent ainsi une localisation en élévation mais aussi en azimut pour les hautes fréquences (supérieures à 4 kHz), précisément là où l'ITD et l'ILD ne permettent plus une localisation fiable.

HRTF

Les HRTF (*Head Related Transfer Function*) sont assimilables à des filtres acoustiques qui décrivent la propagation acoustique entre une source sonore et les oreilles d'un auditeur. Ces filtres dépendent de la position de la source sonore et traduisent l'ensemble des phénomènes de propagation des ondes acoustiques entre la source et l'entrée des conduits auditifs :

- la propagation en champ libre,
- la diffraction par la tête de l'auditeur [Duda and 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 de l'oreille [Batteau, 1967, Shaw and Teranishi, 1968, Hebrank and Wright, 1974].

Les HRTF sont ainsi dépendantes de la morphologie de l'auditeur.

La définition physique des HRTF gauche et droite, H_L et H_R, est donnée par l'équation

$$H_{L,R}(r,\theta,\phi)(j\omega) = \frac{\phi_{L,R}(r,\theta,\phi)(j\omega)}{\phi_0(j\omega)}, \tag{1.2}$$

où (r,θ,ϕ) désigne la position en coordonnées sphériques de la source sonore 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. La figure 1.2 présente un exemple d'HRTF gauche et droite mesurées sur un individu pour une position de source donnée.

Les HRTF rassemblent sous une forme compacte l'ensemble des indices mis à disposition du système auditif pour localiser les sons (ITD, ILD, et indices spectraux). L'ITD peut être calculée à partir de la différence des retards des HRTF droite et gauche de même que l'ILD s'exprime comme la différence entre les spectres d'énergies des HRTF droite et gauche. Pour ce qui est de l'identification des indices spectraux, deux théories s'affrontent. L'une considère que la localisation se base sur l'intégration du spectre global des HRTF par le système auditif

[Middlebrooks, 1992], alors que l'autre théorie accorde plus d'importance aux caractéristiques locales du spectre des HRTF (succession de résonances et d'antirésonances) [Blauert, 1970]. En l'état actuel de la connaissance, il paraît délicat d'écarter ou de valider l'une de ces théories.

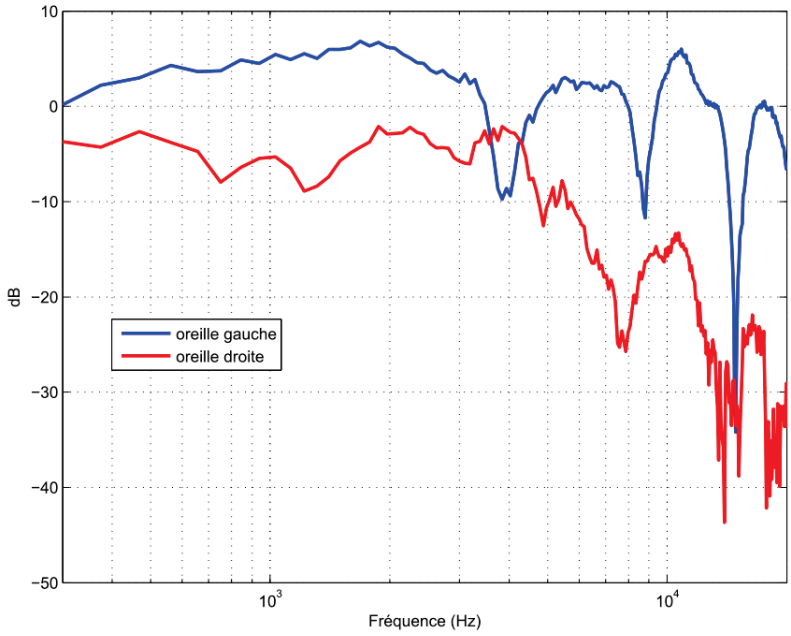


FIGURE 1.2 – *Exemple de HRTF gauche (bleue) et droite (rouge) mesurées sur un individu, tiré de [Nicol, 2010] (base de HRTF JM Pernaux - Orange Labs, direction* $(\phi, \theta) = (-141°, -11°)$ *en coordonnées sphériques).*

1.2.2 Perception de la distance d'une source sonore

Les mécanismes perceptifs utilisés pour la perception de la distance d'une source sonore diffèrent de ceux utilisés pour la perception de la direction. Il est communément admis que la perception de la distance de sources statiques dépend principalement de quatre indices acoustiques : le niveau sonore, la réverbération, le spectre, et les différences interaurales [Mershon and King, 1975, Blauert, 1997, Loomis et al., 1998, Zahorik et al., 2005]. Le niveau sonore et la réverbération sont probablement les deux indices ayant le plus d'impact sur la perception de la distance [Shinn-Cunningham, 2000a].

Niveau sonore

Le niveau sonore est considéré comme le facteur ayant le plus d'impact sur la perception de la distance, car la variation de niveau sonore est aisément détectée par un auditeur. Lorsqu'une source acoustique s'éloigne, son niveau sonore décroît. Dans le cas idéal d'une source sonore ponctuelle placée en champ libre (ou dans un environnement anéchoïque), le niveau sonore observe une décroissance de 6 dB par doublement de la distance r [Bruneau, 1983]. En effet, la pression de l'onde acoustique est proportionnelle à $1/r^2$. Il est souvent mentionné que le niveau sonore est un indice relatif. En effet, il ne permet pas, sans connaissance *a priori* de la

source émettrice, d'estimer la distance absolue à laquelle se situe la source.

Réverbération

Contrairement à l'intensité sonore, le rapport champ direct sur champ réverbéré permet de fournir une information absolue sur la distance d'un objet sonore. Dans un environnement anéchoïque, le niveau sonore d'une source acoustique ponctuelle décroît de 6 dB par doublement de distance. Cette décroissance est modifiée dès lors que la source acoustique est placée dans un environnement échoïque. Dans ce cas, en effet, le champ direct est prédominant par rapport au champ réverbéré à proximité de la source, alors que c'est le champ réverbéré qui devient prédominant lorsqu'on s'éloigne de la source.

Contenu spectral

Le troisième indice acoustique concerne les propriétés spectrales d'une source sonore. Du fait de l'absorption par l'air, les composantes hautes fréquences d'une source acoustique vont être atténuées progressivement avec la distance de propagation. Il a été montré que cet effet peut avoir une influence pour des distances supérieures à 15 mètres [Blauert, 1997]. Comme dans le cas du niveau sonore, cet indice est porteur d'informations si le spectre de la source émettrice est connu. Il s'agit donc d'un indice de localisation relatif.

Différences interaurales

Les différences binaurales constituent le quatrième indice acoustique capable de fournir une information sur la distance d'une source sonore statique, notamment pour les sources situées en champ proche (moins d'un mètre environ) [Shinn-Cunningham et al., 2005, Brungart and Rabinowitz, 1999]. À faible distance, la présence de la tête de l'auditeur représente un obstacle du point de vue de l'onde acoustique, ce qui implique une modification des différences interaurales et plus particulièrement de l'ILD [Duda and Martens, 1998, Brungart and Rabinowitz, 1999].

Les indices acoustiques dynamiques

Lorsque la source sonore (ou bien l'auditeur) est en mouvement, des indices acoustiques dynamiques s'ajoutent aux indices précédemment évoqués. Prenons pour exemple le cas d'un véhicule qui se rapproche d'un auditeur. Pour commencer, ce rapprochement induit une variation du niveau sonore. Si la source sonore se déplace à vitesse constante, l'auditeur va convertir cette variation du niveau sonore en une distance, grâce à l'estimation du *temps avant contact*. Ce temps est également appelé τ acoustique [Ashmead et al., 1995, Shaw et al., 1991]. Ensuite, le rapprochement du véhicule va induire une variation du spectre perçu par l'auditeur : il s'agit de l'effet *Doppler*. Il a été montré que les auditeurs sont toutefois moins sensibles à ce changement fréquentiel qu'à la variation d'intensité causée par le déplacement [Rosenblum et al., 1987]. L'effet de parallaxe acoustique est un autre indice dynamique qui permet d'estimer la distance d'une source sonore. Un déplacement latéral entre la source et l'auditeur induit en effet un changement angulaire plus ou moins important suivant la distance de la source.

Les indices non-acoustiques

Conjointement aux indices acoustiques, d'autres indices peuvent être utiles pour estimer la distance d'une source sonore. La vision, par exemple, peut évidemment donner des informations précieuses sur la localisation d'un objet, mais elle peut également biaiser la perception sonore. Cet effet a été mis en évidence pour des tâches de localisation en azimut et porte le nom d'effet ventriloque [Howard and Templeton, 1966]. Des études mentionnent des effets similaires pour la localisation en distance [Gardner, 1968, Mershon et al., 1980, Bowen et al., 2011, Côté et al., 2011, Hládek et al., 2013]. Ces phénomènes d'interaction audio-visuelle seront développés dans la section 1.4. D'autres études ont montré que la perception de la distance peut également être influencée par la familiarité avec le stimulus sonore employé. D'après ces études, l'utilisation de la parole [Coleman, 1962] ou le fait de multiplier les expériences [Kopčo et al., 2004, Shinn-Cunningham, 2000b] pourraient améliorer la précision des jugements de distance.

1.2.3 Performances du système auditif

Performances de localisation en azimut

La localisation auditive en azimut est plus précise dans la zone frontale avec un flou de localisation (ou *localization blur*) d'environ 3° d'après la littérature [Blauert, 1997]. Des écarts allant de 1° à 5° sont rapportés en fonction de la nature des stimuli employés dans les différentes expériences. Cette précision se dégrade progressivement lorsque les sources se rapprochent de l'axe interaural. La localisation dans le plan horizontal est moins discriminante dans le cas de sources latérales (situées sur l'axe interaural) avec une précision d'environ 10°.

Le pouvoir de discrimination du système auditif pour la localisation dans le plan horizontal peut être caractérisé par le MAA (*Minimum Audible Angle*) [Mills, 1958]. Le MAA est défini comme étant la plus petite différence angulaire perceptible par l'auditeur (JND, *Just Noticeable Difference*), lorsque deux sons successifs sont présentés. Dans des conditions idéales, la plupart des auditeurs sont capables de détecter des changements angulaires de 1° lorsque la source est placée dans la zone frontale. Dans le cas d'un son pur (sinusoïde), cette précision diminue pour les fréquences comprises entre $1,5\ kHz$ et $2\ kHz$, comme l'illustre la figure 1.3. Suivant la fréquence, les valeurs de MAA varient entre 1° et 3° pour une source placée à 0° par exemple.

Performances de localisation en élévation

Les performances de localisation auditive en élévation sont moins bonnes qu'en azimut. Blauert mentionne des performances différentes suivant le stimulus utilisé [Blauert, 1997] : pour une source placée dans la zone frontale, le flou de localisation est d'environ 4° pour un bruit blanc [Wettschurek, 1971], 9° pour une voix familière [Damaske and Wagener, 1969], et 17° pour la voix d'une personne inconnue [Blauert, 1970]. La précision de localisation auditive en élévation varie également en fonction de l'élévation. Blauert reporte des écarts allant de 9° pour une source placée en zone frontale, à 22° pour une source placée au dessus de la tête de l'auditeur. L'erreur de localisation en élévation est donc minimale en zone frontale.

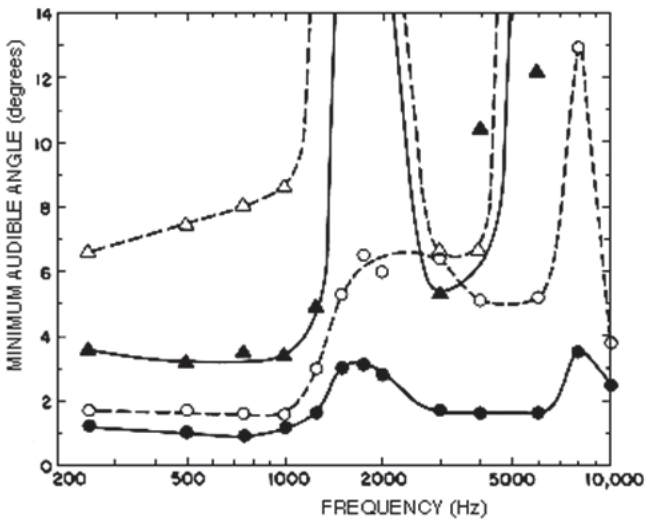


FIGURE 1.3 – *Évolution de l'angle minimum audible (MAA) en fonction de la fréquence du son pur utilisé et de l'azimut de la source (*0° *cercle noir,* 30° *cercle blanc,* 60° *triangle noir,* 75° *triangle blanc), tiré de [Mills, 1958].*

Performances de localisation en distance

Globalement, les performances de localisation auditive en distance semblent assez médiocres. Zahorik montre un phénomène de compression entre la distance réelle d'une source sonore et la distance perçue du stimulus sonore [Zahorik, 2001, Zahorik, 2002a]. Pour les distances inférieures à environ 1,5 mètre, les distances sont systématiquement surestimées alors qu'elles sont sous-estimées au-delà. Zahorik remarque que la loi de puissance de Stevens [Stevens, 1957] est une bonne approximation de la relation entre la distance perçue et la distance physique des sources sonores. Cette loi s'écrit sous la forme de la fonction

$$d_p = k.{d_r}^a, \tag{1.3}$$

où d_p est la distance perçue, d_r est la distance réelle, k et a sont les paramètres d'ajustement de la fonction puissance. Zahorik analyse les résultats issus de 84 études et trouve que la valeur moyenne de k est légèrement supérieure à 1, et que la valeur moyenne de l'exposant a est égale à $0,54$ [Zahorik et al., 2005]. Cet exposant traduit la compression de la perception de la distance ($a < 1$).

La variabilité des jugements de distance est relativement importante puisqu'elle atteint 20 à 60 % de la distance réelle de la cible suivant les participants [Zahorik, 2002b]. Blauert [Blauert, 1997] mentionne l'importance de la familiarité avec la source sonore dans la précision de la localisation en distance. Par exemple, les performances de localisation en distance seraient meilleures pour une voix familière que pour des sources moins familières. De même, le fait de faire varier l'élocution peut entraîner une surestimation (si on utilise des cris comme stimulus auditif) ou sous-estimation (avec des chuchotements) de la distance perçue [Gardner, 1969].

1.3 Perception visuelle de l'espace

Le champ de vision monoculaire s'étend dans le plan horizontal de 100° côté tempe à 60° côté nez, ce qui crée une zone de recouvrement d'environ 120° dans laquelle la vision est binoculaire [Spector, 1990]. Le champ de vision vertical est quant à lui limité à 40° vers le haut et 70° vers le bas. La figure 1.4 illustre le champ visuel monoculaire et binoculaire humain.

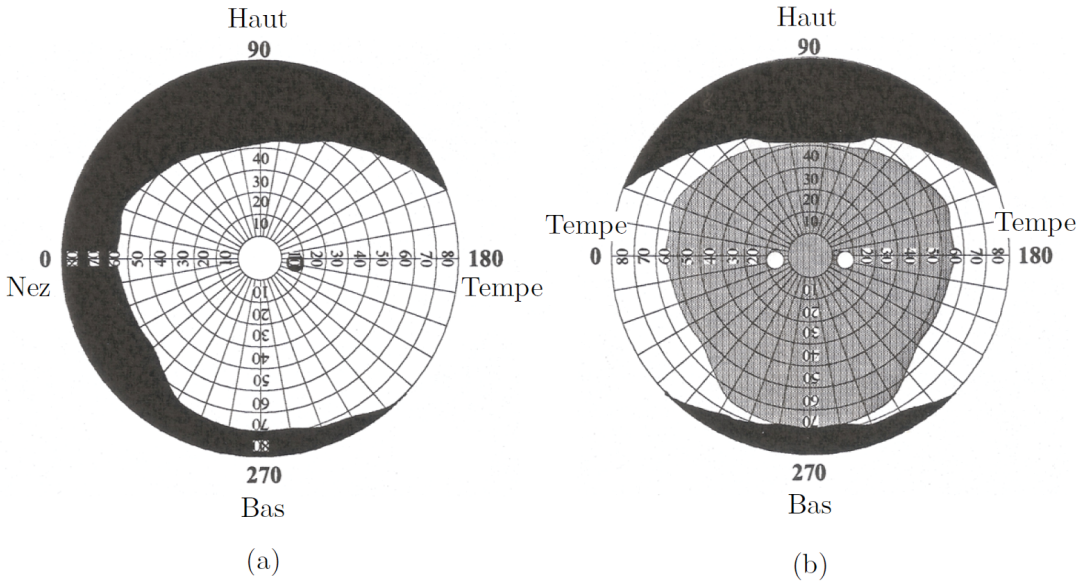


FIGURE 1.4 – *Champ visuel monoculaire droit (a) et binoculaire (b), tiré de [Fournier, 1995], d'après [Ruch and Fulton, 1960]. Les zones blanches représentent les plages de vision monoculaire, la région grisée indique la zone de vision binoculaire.*

1.3.1 Perception de la profondeur visuelle

Comme pour le système auditif, qui utilise des informations monaurales, binaurales et dynamiques, de nombreux indices visuels permettent à l'être humain d'évaluer la localisation en profondeur d'objets visuels [Cutting and Vishton, 1995, Palmer, 1999] : des indices monoculaires, binoculaires, mais aussi ceux liés aux mouvements de l'objet (ou de l'observateur).

Indices perceptifs monoculaires

— L'**occlusion** permet de positionner relativement les objets en profondeur. En effet, si un objet masque un autre objet, il est alors facile pour l'observateur de savoir quel objet est placé devant l'autre (figure 1.5 (a)). L'occlusion est donc un indice de distance relative.

— La **hauteur dans le champ visuel** d'un objet par rapport à la ligne d'horizon permet d'estimer la distance de manière absolue. En effet, plus un objet est éloigné, plus il se rapproche de la ligne d'horizon.

— La **taille relative des objets** donne des informations sur la distance absolue d'un objet dès lors que ses dimensions sont connues par l'observateur. Dans le cas contraire, cet indice monoculaire ne permet que de réaliser des jugements relatifs de distance (figure 1.5 (b)). Plus la projection rétinienne d'un objet visuel est grande, plus l'objet est perçu comme étant proche.
— La **perspective linéaire** est souvent utilisée en peinture pour créer une impression de profondeur (lignes au sol, orientation des murs). La perspective linéaire impose que deux lignes parallèles convergent vers un point de fuite (figure 1.5 (c)). La présence de cet indice dans une image augmente la sensation de relief visuel.
— La **perspective aérienne** est un indice monoculaire qui permet d'estimer la position absolue d'un objet, pour des distances importantes. L'humidité et la pollution atmosphérique peuvent en effet diminuer le contraste ou la visibilité d'un objet.
— Les **ombres**, tout comme les variations de lumière, donnent à un objet de l'épaisseur et créent ainsi une sensation de relief. La figure 1.5 (d) illustre le relief apporté par une ombre (dégradé de gris).
— Le **gradient de texture** varie également en fonction de la distance d'un objet visuel. Plus un objet est proche, plus il sera facile de distinguer la texture de sa surface.
— L'**accommodation** résulte de la déformation du cristallin des yeux. Ce phénomène adaptatif permet à l'œil d'effectuer la mise au point sur des objets proches ou lointains et assure la netteté des images projetées sur la rétine. Cet indice monoculaire permet une estimation absolue de la distance d'un objet visuel.

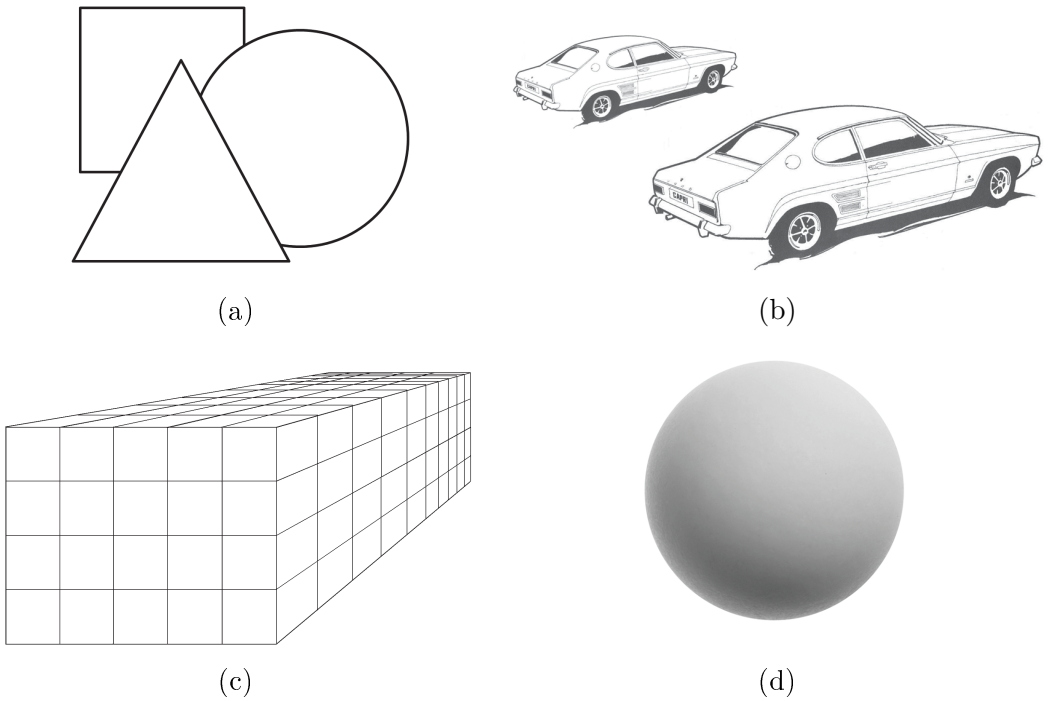


FIGURE 1.5 – *Illustration d'indices monoculaires disponibles pour estimer la distance d'objets visuels : (a) occlusion, (b), taille relative des objets, (c) perspective linéaire, (d) effet d'ombrage.*

Indices perceptifs binoculaires

— La **disparité binoculaire** est la différence entre les projections rétiniennes des deux yeux lorsqu'un observateur fixe un objet. Cette différence de point de vue est causée par le décalage entre les yeux d'un observateur. Il a été montré que l'écart interoculaire est en moyenne de 62 *mm* chez la femme et 65 *mm* chez l'homme adulte [Dodgson, 2004]. La disparité binoculaire permet de percevoir le relief et donc de définir la position relative des objets devant et derrière le point de fixation.
L'horoptère théorique, également appelé cercle de *Vieth Müller*, est le cercle qui passe par le point de fixation et l'ensemble des points ayant des images strictement identiques sur les deux rétines (points présentant une disparité nulle). Les points qui ne sont pas placés sur l'horoptère présentent des disparités binoculaires. Tant que ces disparités sont faibles, les deux points de vue sont fusionnés par notre cerveau en un simple percept. La zone autour de l'horoptère dans laquelle la fusion binoculaire a lieu est appelée aire de Panum (figure 1.6). En dehors de cette zone, les deux images ne peuvent être fusionnées, ce qui engendre une vision double (phénomène de diplopie).

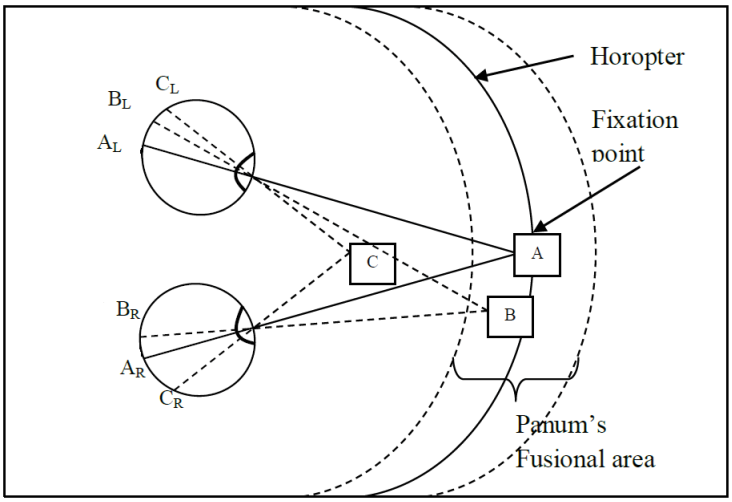


FIGURE 1.6 – *Horoptère et aire de Panum, d'après [Patterson, 2007]. Les points A_R et A_L correspondent respectivement aux projections du point de fixation A sur les rétines droite et gauche. L'objet B est placé à l'intérieur de la zone de Panum et est vu en simple par l'observateur, alors que l'objet C engendre une disparité binoculaire trop importante pour être fusionnée par le cerveau. L'objet C est donc vu en double (phénomène de diplopie).*

— La **convergence** des axes optiques des yeux est liée au point de fixation. Plus le point de fixation est proche, plus la convergence des yeux est importante. Cet indice permet ainsi une estimation à la fois absolue et relative de la distance d'objets visuels.

Indices perceptifs dynamiques

— La **parallaxe du mouvement** peut être interprétée, de manière absolue ou relative, comme une distance. En effet, comme pour la modalité auditive, un déplacement latéral entre l'objet et l'observateur induit un changement angulaire (de la projection

rétinienne, cette fois-ci) qui va directement dépendre de la distance de l'objet.
— Le τ **visuel** définit le temps avant l'impact entre un objet et l'observateur, si l'un des deux se déplace à vitesse constante.

Importance relative des indices perceptifs visuels

La perception de la distance d'un objet visuel dépend donc d'un nombre important d'indices monoculaires et binoculaires, dans des conditions statiques ou dynamiques. Cependant, le cerveau va accorder plus ou moins d'importance à ces indices, en fonction de la distance des objets visuels. D'après E. Cutting et M. Vishton, l'espace visuel peut être divisé en trois zones : l'espace personnel (pour les distances inférieures à 2 mètres), l'espace d'action (pour les distances comprises entre 2 et 20 mètres), et l'espace lointain (pour les distances supérieures à 20 mètres) [Cutting and Vishton, 1995]. Cutting et Vishton avancent que l'importance relative des indices mentionnés précédemment varie en fonction de ces zones (figure 1.7).

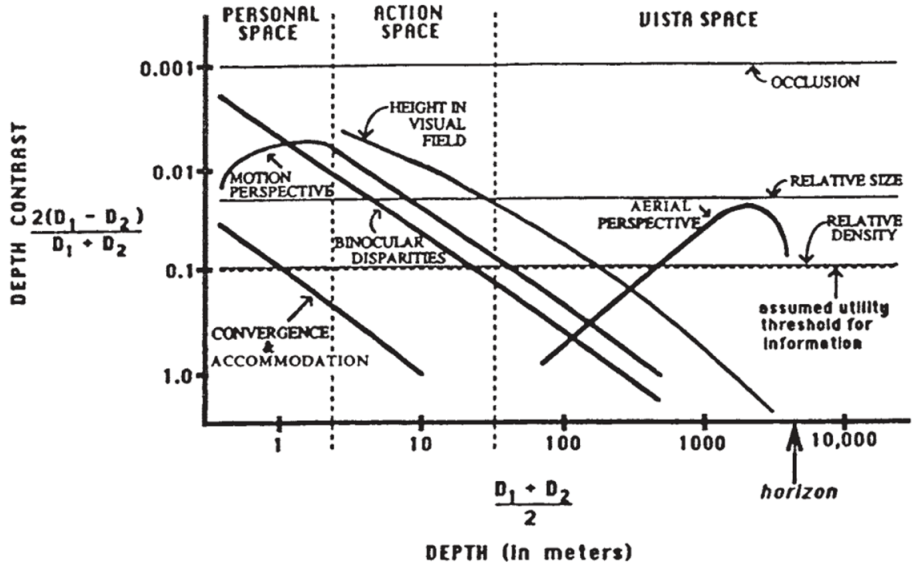


FIGURE 1.7 – *Importance relative des différents indices dans l'évaluation de la profondeur visuelle, d'après [Cutting and Vishton, 1995].*

1.3.2 Performances du système visuel

L'acuité visuelle monoculaire est très précise dans le champ de vision central (un cône de 2°) et se dégrade à mesure que l'objet sort de cette zone (champ de vision périphérique). Dans le cône de champ visuel central, la valeur minimum de l'angle sous lequel deux points sont vus séparément est de l'ordre d'une minute d'arc (soit 1/60 de degré) [Howard, 1982]. Ce seuil de discrimination entre deux objets est de l'ordre de 2° pour une cible visuelle placée à un azimut de 25° [Perrott, 1993].

L'acuité visuelle binoculaire définit la capacité de l'être humain à distinguer une différence de profondeur entre deux objets. Cette acuité est maximale dans la région de l'horoptère

et décroît lorsque la cible s'éloigne de cette zone. Une étude a illustré les différences inter-individuelles en mesurant l'acuité stéréoscopique d'environ 200 personnes [Coutant and Westheimer, 1993]. Les résultats de cette étude montrent que 97 % des participants détectent une différence de profondeur pour une disparité binoculaire de 2 minutes d'arc, et que ce seuil atteint 30 arcsecondes pour 80 % des sujets. La disparité minimale détectable par l'œil est de l'ordre de 2 à 5 arcsecondes pour les meilleurs sujets. Une acuité stéréoscopique de 2 arcsecondes correspond à la capacité de détecter une différence de profondeur de 4 mm à une distance de 5 mètres.

1.4 Perception de l'environnement audio-visuel

Nous avons abordé dans les paragraphes précédents les mécanismes perceptifs permettant la localisation de stimuli sonores ou visuels. Toutefois, les stimuli présents dans notre environnement ne sont généralement pas unimodaux mais multimodaux. Ces stimulations multisensorielles doivent donc être intégrées simultanément pour être interprétées de manière compréhensible comme un percept unique. Notre cerveau, et notamment le colliculus supérieur, permet d'intégrer des informations visuelles et auditives lorsqu'elles sont coïncidentes (temporellement et spatialement). Cependant, la plasticité de notre cerveau est telle que des décalages entre les stimuli auditifs et visuels peuvent être tolérés (décalage spatial ou désynchronisation).

1.4.1 Interactions et illusions audio-visuelles

Impact de la présentation bimodale sur la localisation

Une conséquence bien connue de la présentation audio-visuelle sur les tâches de localisation est le phénomène de **capture visuelle** ou **effet ventriloque** [Howard and Templeton, 1966, Thurlow and Jack, 1973]. L'effet ventriloque se produit lorsque des stimuli audio et visuel sont temporellement coïncidents mais spatialement disparates. Ce décalage spatial entre les deux stimulations a généralement pour effet de biaiser le jugement de localisation du stimulus sonore. En d'autres termes, la position du stimulus sonore est attirée par le stimulus visuel associé. Ce biais perceptif est d'ailleurs largement utilisé dans les applications, telles que le cinéma ou la télévision, où les spectateurs ont l'impression que les voix proviennent de la bouche des acteurs alors que les haut-parleurs sont placés de chaque côté de l'écran. Nous aborderons les bornes spatiales entre lesquelles l'effet ventriloque se produit à la fin de ce chapitre.

La localisation d'un objet visuel peut également être biaisée par le stimulus sonore si la localisation par la modalité visuelle est ambiguë (par exemple si l'objet visuel est flou). On parle alors d'**effet ventriloque inverse** [Alais and Burr, 2004]. Cette dominance de l'audition sur la vision dans une tâche de localisation est toutefois rare.

Illusions perceptives audio-visuelles

L'**effet McGurk** [McGurk and MacDonald, 1976] est une illusion perceptive dans laquelle les stimulations auditives et visuelles donnent naissance à un nouveau percept audio-visuel. Dans cette expérience, le participant observe une vidéo dans laquelle une personne prononce la syllabe /ga/ alors que le son /ba/ est diffusé. Cette incohérence entre les stimuli sonore et

visuel donne l'illusion que la syllabe /da/ a été présentée. La création de ce percept pourrait s'apparenter à la réponse la plus plausible face à cette ambiguïté. En effet, la prononciation de la consonne /g/ implique une occlusion du conduit vocal au niveau de la base de la langue, alors que la consonne /b/ est labiale (l'occlusion du conduit vocal est réalisée au niveau des lèvres). Le percept /d/ qui résulte de cette stimulation audio-visuelle incohérente est, quant à elle, une consonne dentale, et l'occlusion du conduit vocal se fait au niveau des dents (entre les lèvres et la base de la langue).

L'**illusion du rebond** [Sekuler et al., 1997] est également le résultat de l'ambiguïté du stimulus visuel : deux objets visuels en mouvement suivent une trajectoire convergente. Cependant, il est impossible de savoir si les objets se croisent en passant l'un sur l'autre, ou s'ils se heurtent et rebondissent. Le fait d'avoir un son de choc présenté au moment du croisement des éléments visuels ou deux sons présentés en continu influence la compréhension de la scène visuelle. Lorsque le son de choc est présenté au moment du croisement des objets, plus de sujets estiment que les objets rebondissent.

La **capture auditive** traduit la dominance de la modalité auditive sur la modalité visuelle dans le domaine temporel. L'illusion du *double-flash* est un exemple de capture auditive [Shams et al., 2000]. Il s'agit de présenter simultanément un flash lumineux et deux sons brefs. Les participants ont alors l'impression d'avoir vu deux flashs lumineux. Fendrich and Corballis [Fendrich and Corballis, 2001] ont, quant à eux, mis en évidence un effet ventriloque temporel : le moment d'apparition du stimulus visuel est biaisé par la présence du son qui est émis avant ou après le stimulus visuel. D'autres exemples mettent en évidence l'influence de l'audition sur la fréquence d'oscillation perçue d'un stimulus audio-visuel [Shipley, 1964], ou sur la durée d'apparition du stimulus [Walker and Scott, 1981].

D'après Welch et Warren [Welch and Warren, 1980], le phénomène de capture (visuelle ou auditive) s'expliquerait par l'hypothèse de la modalité pertinente. Cette hypothèse consiste à dire que, du fait des différences de sensibilité des systèmes perceptifs, la vision domine l'audition pour les tâches de localisation, et que l'audition domine la vision pour les tâches temporelles. Cette hypothèse simplificatrice est cependant controversée : Alais, Newell, et Mamassian mettent en avant le modèle du maximum de vraisemblance comme étant plus représentatif des phénomènes de perception multimodale [Alais et al., 2010].

1.4.2 Modèle d'intégration multimodale

Le modèle du maximum de vraisemblance permet de prédire la perception bimodale ou multimodale en prenant en compte la fiabilité des informations unimodales disponibles [Ernst and Banks, 2002, Alais and Burr, 2004]. Dans le cas de la localisation spatiale, l'être humain est capable de localiser un stimulus visuel situé à une position m_V avec une variance ${\sigma_V}^2$, et un stimulus audio situé à une position m_A avec une variance ${\sigma_A}^2$. D'après le modèle du maximum de vraisemblance, un stimulus audio-visuel est alors localisé autour d'une position m_{AV} avec une variance ${\sigma_{AV}}^2$ données par :

$$m_{AV} = \frac{{\sigma_V}^2}{{\sigma_V}^2 + {\sigma_A}^2} m_A + \frac{{\sigma_A}^2}{{\sigma_V}^2 + {\sigma_A}^2} m_V, \tag{1.4}$$

et

$$\sigma_{AV}{}^2 = \frac{\sigma_A{}^2 \sigma_V{}^2}{\sigma_A{}^2 + \sigma_V{}^2} \leq \min\left(\sigma_A{}^2, \sigma_V{}^2\right). \tag{1.5}$$

Cette définition permet ainsi de pondérer les modalités (ou les stimulations) proportionnellement à leur fiabilité. La prise en compte de la saillance des informations disponibles permet par exemple de prédire à la fois l'effet ventriloque et l'effet ventriloque inverse contrairement à l'hypothèse de la modalité pertinente. De plus, le modèle du maximum de vraisemblance prend en compte la contribution des différentes modalités, même si l'une d'elles est dominante.

1.4.3 Limites spatio-temporelles de l'intégration audio-visuelle

Les phénomènes de fusion perceptive ou d'intégration multi-sensorielle décrits précédemment ont pour conséquence la compréhension d'un percept unique multidimensionnel. Dans le cas de la perception audio-visuelle, la plasticité du cerveau permet de maintenir cette perception unique lorsque des décalages (temporels ou spatiaux) sont introduits. En effet, tant que ces décalages sont faibles, la fusion perceptive se produit. Cependant, ce phénomène d'intégration est rompu dès lors que les écarts introduits entre les stimuli sonores et visuels sont trop importants. Cette partie a pour but de définir les écarts temporels et spatiaux généralement tolérés par notre système perceptif, garantissant l'intégration audio-visuelle.

Fenêtre d'intégration temporelle

La fenêtre d'intégration audio-visuelle temporelle a fait l'objet d'études dans le domaine de la psychophysique. Il a été montré que le point de synchronie subjective (PSS) entre un flux sonore et un flux visuel est variable suivant les individus, mais qu'il est différent de la synchronie physique [Stone et al., 2001]. En d'autres termes, un décalage temporel entre le son et l'image est jugé comme étant plus synchrone qu'une stimulation audio-visuelle ne présentant pas de décalage. La simultanéité audio-visuelle semble en effet détectée lorsque le son présente un retard d'environ 30 à 50 ms [Stone et al., 2001, Spence et al., 2003, Kohlrausch and van de Par, 2005]. Ce phénomène intrigant s'explique par la vitesse de propagation des ondes qui est différente entre le son ($340\ m.s^{-1}$) et la lumière ($3.10^8\ m.s^{-1}$). Cette différence impose que, dans le monde physique qui nous entoure, un son arrive toujours après l'évènement visuel associé. Ce décalage temporel se retrouve d'ailleurs sur la zone de synchronie avec des seuils de détection compris entre $-40\ ms$ (son en avance) et $+100\ ms$ (son en retard) pour des stimuli impulsionnels ou de parole [Hollier and Rimell, 1998] (voir figure 1.8).

D'autres études ont montré que le stimulus peut avoir une influence sur l'intégration audio-visuelle, avec des fenêtres d'intégration temporelle allant de $-75\ ms$ à $+188\ ms$ pour un bruit impulsif associé à la vidéo d'un impact de marteau, et $-131\ ms$ à $+258\ ms$ pour un locuteur [Dixon and Spitz, 1980]. Les différents stimuli et protocoles utilisés dans ces études expliquent probablement les différences de seuil constatées. Cependant, toutes ces études s'accordent sur le fait que la fenêtre d'intégration audio-visuelle temporelle est asymétrique, et que la présence d'asynchronie est mieux tolérée lorsque le son est en retard sur l'image que la situation inverse.

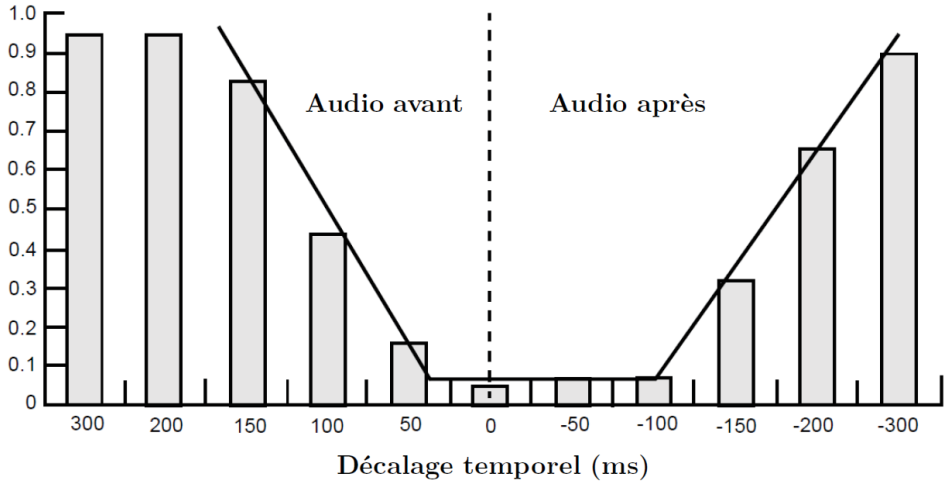


FIGURE 1.8 – *Plateau de perceptibilité d'asynchronie audio-visuelle, d'après [Hollier and Rimell, 1998].*

Fenêtre d'intégration spatiale en azimut

De nombreuses études ont visé à déterminer l'écart maximum tolérable lorsqu'un stimulus sonore et un stimulus visuel ne sont pas positionnés au même endroit (différence de localisation en azimut). Pour un stimulus audio-visuel abstrait, représentant par exemple un spot lumineux associé à un sinus pur, il semblerait que la fenêtre d'intégration audio-visuelle soit d'environ 3° [Lewald et al., 2001].

Différentes études se sont intéressées à des stimuli plus réalistes, impliquant par exemple un visage animé et un signal de parole. Il semblerait que, dans ce cas, notre système perceptif est davantage tolérant. Komiyama s'est par exemple intéressé à la gêne perçue suite à l'introduction d'une différence de localisation audio-visuelle dans des applications telles que la télévision [Komiyama, 1989]. D'après Komiyama, l'écart angulaire à ne pas dépasser est défini par le niveau de gêne entre une dégradation "perceptible mais non gênante", et une dégradation "légèrement gênante". Compte tenu de cette définition, l'écart angulaire admissible est de 20° pour des sujets naïfs. Environ 50 % des participants ont été capables de détecter un décalage spatial de 10° alors que cet écart angulaire peut atteindre 30° avant d'être jugé comme "légèrement gênant" par plus de la moitié des sujets. Une étude comparable a permis de déterminer les bornes de la fenêtre d'intégration grâce à un protocole expérimental inspiré de la psychophysique [Munoz Soto et al., 2008]. Dans cette expérience, les participants indiquent simplement si oui ou non les stimuli audio-visuels présentés sont spatialement cohérents. L'utilisation de ce paradigme de type "oui / non" permet de construire une courbe psychométrique. Les auteurs considèrent la limite de la fenêtre d'intégration comme la valeur au-delà de laquelle plus de 50 % des participants détectent une incohérence. Les résultats de cette étude montrent qu'un écart angulaire de 15° est tolérable dans le cas d'un visage animé associé à une voix. De la même manière, des écarts angulaires de l'ordre de 18° semblent être tolérés dans le cas d'objets audio-visuels virtuels [André et al., 2014], où un signal de parole diffusé par *Wave Field Synthesis* accompagne un personnage virtuel présenté en 3D.

Fenêtre d'intégration spatiale en élévation

En comparaison du nombre important de travaux menés sur l'intégration audio-visuelle dans le plan horizontal, peu d'études se sont intéressées aux limites d'intégration audio-visuelle en élévation. De manière générale, il semble que l'intégration se fasse plus facilement en élévation qu'en azimut [Godfroy et al., 2003, Hartnagel, 2007], du moins dans le plan médian. En effet, les écarts angulaires tolérables en élévation sont environ deux fois plus importants qu'en azimut, avec des valeurs moyennes proches de 11° en azimut contre $19,5°$ en élévation sur l'ensemble des conditions testées par Hartagnel [Hartnagel, 2007] (stimuli audio-visuels présentant des écarts angulaires allant jusqu'à 30° en azimut et 20° en élévation). Cette étude a également montré que l'étendue des aires de fusion varie en fonction de la position des sources en azimut (augmentation de l'étendue avec l'excentricité) mais pas en élévation. Ces résultats font écho à une étude antérieure [Godfroy et al., 2003] dans laquelle des zones d'intégration d'environ 13° en azimut et $21,5°$ en élévation sont obtenues (stimulus auditif : salve de 500 ms de bruit rose, stimulus visuel : spot lumineux). La figure 1.9 illustre les aires de fusion déterminées dans cette étude. Il est important de noter que les valeurs avancées ici correspondent à l'écart angulaire à partir duquel 50 % des participants estiment que les stimuli sonores et visuels sont spatialement incohérents.

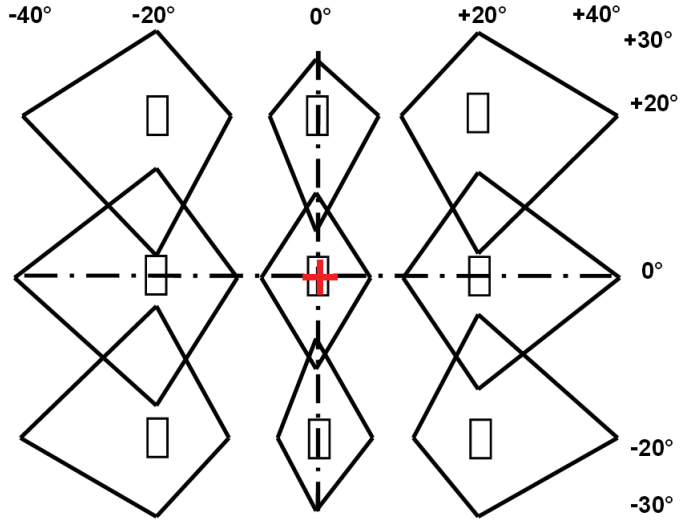


FIGURE 1.9 – *Étendue des zones de fusion audio-visuelles en fonction de la position des haut-parleurs (représentés pas les rectangles), d'après [Godfroy et al., 2003]. La croix rouge placée au centre représente le point de fixation visuel.*

Fenêtre d'intégration spatiale en distance

Dans le cas de stimuli brefs, l'intégration audio-visuelle en distance serait principalement basée sur la différence de vitesse de propagation entre les ondes acoustiques et les ondes lumineuses [Sugita and Suzuki, 2003, Alais and Carlile, 2005, Burr and Alais, 2006]. Burr et Alais ont prouvé l'existence d'une relation linéaire entre la distance séparant le sujet de la source,

et le décalage temporel correspondant à la sensation de simultanéité audio-visuelle perçue. D'après Burr et Alais, la pente de cette relation linéaire est de l'ordre de 3 $ms.m^{-1}$ (les différences inter-individuelles observées vont de 2,5 à 4,2 $ms.m^{-1}$ avec une valeur moyenne de 3,2 $ms.m^{-1}$). Cette valeur semble correspondre au délai nécessaire pour compenser la différence entre les vitesses de propagation du son et de la lumière (délai estimé à 2,9 $ms.m^{-1}$ à 20°C). Cette hypothèse reste cependant controversée [Arnold et al., 2005, Heron et al., 2007]. Heron et al., par exemple, ont montré que le cerveau humain ne serait pas capable d'estimer le décalage plausible entre le temps d'arrivée du stimulus sonore et l'élément visuel, et de décider de la simultanéité des stimuli.

Des études ont récemment été menées dans le but d'estimer la zone d'intégration audio-visuelle en distance dans le cas de stimuli réalistes, en utilisant des objets audio-visuels virtuels [Gorzel et al., 2012, Corrigan et al., 2013]. Dans l'étude réalisée par Gorzel *et al.*, le stimulus visuel est un haut-parleur présenté de manière aléatoire à 2, 4, ou 8 mètres grâce à un rendu 3D stéréoscopique. Deux stimuli sonores sont utilisés : des pulses de bruit rose, et un signal de parole. Les sources sonores virtuelles sont synthétisées par *Higher Order Ambisonics*, et peuvent être simulées à 11 distances différentes comprises entre 1 et 9 mètres. Le jugement est effectué par un choix forcé, puisque pour chaque présentation audio-visuelle, les participants indiquent s'ils perçoivent le son comme étant placé devant, derrière, ou à l'endroit de l'objet visuel. La zone d'intégration est définie par les auteurs comme les conditions pour lesquelles l'intégration audio-visuelle est perçue par plus de 50 % des participants. Les résultats montrent que plusieurs conditions sonores fournissent une sensation de percept audio-visuel unique. Comme l'illustre

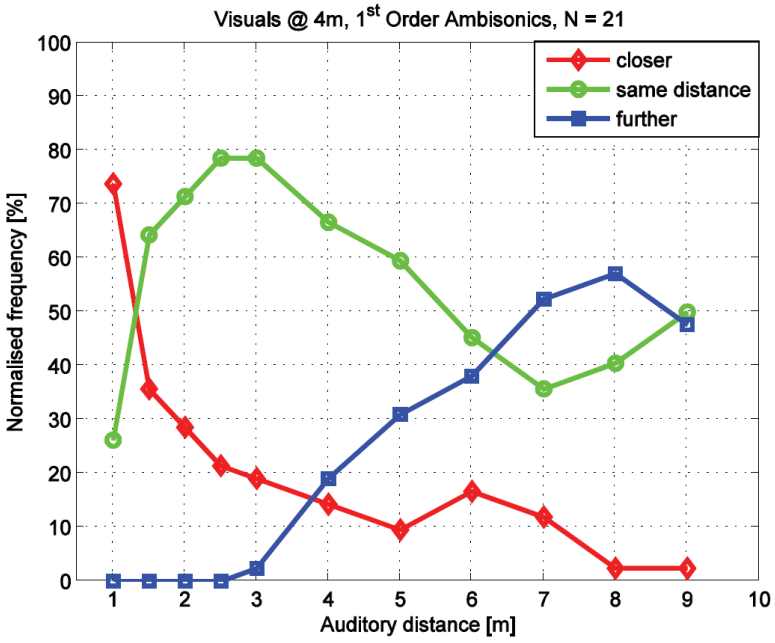


FIGURE 1.10 – *Pourcentage de conditions perçues comme étant spatialement cohérentes ou incohérentes, en fonction de la position des sources sonores virtuelles pour une cible visuelle placée à 4 mètres. Estimations pour le son placé devant (en rouge), derrière (en bleu), ou au même endroit que l'objet visuel (en vert). D'après [Gorzel et al., 2012].*

la figure 1.10, la zone d'intégration pour un objet visuel modélisé à 4 mètres des participants s'étend sur 4,34 mètres (intégration lorsque le son est placé entre 1,31 et 5,65 mètres). L'effet du type de stimulus sonore utilisé n'a pas été mis en évidence dans cette étude. De plus, aucune conclusion ne peut être tirée sur l'évolution de l'étendue des zones d'intégration audio-visuelles en fonction de la distance de l'objet visuel.

1.4.4 Un système perceptif plastique

Il a été montré précédemment que grâce à sa plasticité, notre cerveau est capable sous certaines conditions, d'intégrer des informations visuelles et auditives disparates, et de les interpréter comme un percept multimodal unique. Nous avons choisi de présenter différents phénomènes qui mettent l'accent sur cette capacité d'adaptation et de flexibilité de notre système perceptif. Nous verrons également que de nombreux facteurs cognitifs peuvent aussi avoir un impact sur la perception d'évènements audio-visuels.

Phénomènes d'adaptation

Comme mentionné dans la première partie de ce chapitre, le système auditif permet, au quotidien, de localiser des sources sonores dans l'espace. Il a également été montré que ce système de décodage spatial sonore est propre à chaque individu et que les HRTF rassemblent sous forme compacte tous les indices de localisation sonore. Ce principe de décodage individuel est réalisé par apprentissage tout au long de notre vie. D'après R. Nicol, *"nous construisons par apprentissage le décodeur associé à l'encodeur déterminé par notre morphologie. Ce décodeur présente une certaine flexibilité, afin de s'adapter aux évolutions de notre morphologie (...)."* [Nicol, 2010].

Une étude réalisée par Hofman et al. montre que notre cerveau est capable de s'adapter à un changement morphologique sur une échelle temporelle relativement courte [Hofman et al., 1998]. Dans cette étude, quatre participants ont porté des prothèses pendant six semaines afin de modifier la morphologie de leur pavillon d'oreille. Cette modification morphologique a pour conséquence directe une modification de leur système d'encodage spatial sonore. Un test de localisation réalisé suite à la pose des prothèses montre que les performances de localisation en élévation sont nettement diminuées. Des tests de localisation ont été réalisés tout au long de la période de l'étude. Il apparaît qu'au cours du temps, les sujets se familiarisent avec ce nouveau système d'encodage puisque leurs performances de localisation s'améliorent progressivement. À la fin des six semaines, les prothèses sont retirées et un nouveau test de localisation révèlent que les participants ont gardé leurs performances de localisation initiales (avant la mise en place des prothèses). Ces résultats indiquent que le système auditif s'est adapté à ce nouveau système d'encodage (avec les prothèses), sans pour autant empêcher la compréhension des indices de localisation d'origine (sans les prothèses) qui ont été gardés en mémoire. En ce sens, les auteurs avancent que ce phénomène est comparable à celui de l'apprentissage d'une deuxième langue.

Un autre exemple d'adaptation et de mémorisation de notre système perceptif est le phénomène d'*after-effect* spatial [Lewald, 2002]. Lewald montre que la modalité visuelle peut en effet biaiser la localisation d'un stimulus sonore grâce à un phénomène d'apprentissage. Dans cette expérience, une disparité audio-visuelle de 20° est présentée de manière répétée entre un spot

lumineux et un signal sonore (bruit ou sinus pur). Lewald a montré qu'après cette exposition prolongée à un stimulus bimodal disparate, la localisation d'un objet sonore (unimodal) est systématiquement altérée et attirée vers la localisation du stimulus visuel initialement présent. Ce biais est de l'ordre de $2,4°$ pour le bruit, $3,1°$ pour un sinus pur à 1 kHz, et $5,8°$ pour un sinus pur à 4 kHz.

Des effets de recalibration temporelle du système perceptif audio-visuel ont également été mis en évidence [Fujisaki et al., 2004]. Dans cette étude de Fujisaki *et al.*, un décalage temporel est introduit entre un stimulus sonore (sinus pur) et un stimulus visuel (flash). Les participants sont exposés à ce stimulus audio-visuel pendant 3 minutes. Après cette phase d'adaptation, les sujets doivent estimer le point d'alignement temporel entre un stimulus sonore et un stimulus visuel (jugement de simultanéité). Les résultats montrent que le point de simultanéité subjective est déplacé dans le sens du décalage temporel introduit lors de la phase d'adaptation.

Influence des facteurs cognitifs

Les différents exemples mentionnés dans ce chapitre mettent en évidence l'importance de certains facteurs cognitifs dans la perception audio-visuelle. Il a, en effet, été montré que le lien sémantique existant entre les stimuli sonores et visuels peut avoir une influence sur la manière dont les informations spatiales audio-visuelles sont intégrées. La fenêtre d'intégration audio-visuelle en azimut, par exemple, est plus ou moins large suivant le couple de stimuli employé. Cette fenêtre est de l'ordre de 3° pour des stimuli abstraits tels qu'un spot lumineux associé à un bruit blanc alors qu'un décalage allant jusqu'à 15 à 20° est toléré lorsque le stimulus audio-visuel est un personnage qui parle. Dans ce dernier cas, la relation sémantique entre le son et l'image est importante, ce qui a pour conséquence une tolérance de notre système perceptif plus importante face aux écarts angulaires. L'illusion du rebond [Sekuler et al., 1997], présentée dans la partie 1.4.1, illustre également l'importance du lien sémantique entre stimulus sonore et visuel sur la compréhension d'une scène audio-visuelle ambiguë.

La perception unifiée ou non de stimuli sonores et visuels dépend aussi des attentes du sujet sur la possibilité de percevoir des stimuli issus d'une cause commune. Cette notion est appelée la "présomption d'unité" [Welch, 1999]. Daprès Welch, la présomption d'unité dépend notamment des instructions et de la familiarité des associations multimodales, et aurait pour conséquence la minimisation des divergences intermodales perçues. Il semblerait donc que plus la présomption d'unité est forte, plus l'amplitude du biais intersensoriel est importante. Vatakis et Spence ont montré que la présomption d'unité favorise l'intégration audio-visuelle temporelle dans le cas d'une voix associée à un locuteur [Vatakis and Spence, 2007].

1.5 Conclusion

Le système auditif utilise différents indices monauraux et binauraux pour localiser les sons en azimut et en élévation (ITD, ILD, et indices spectraux). Ces indices sont rassemblés sous forme compacte dans les HRTF qui sont propres à chaque individu et rendent compte de la relation individuelle entre leur spécificité morphologique et leur encodage spatial. La perception de la distance d'une source sonore fait, quant à elle, intervenir des indices acoustiques tels que le niveau sonore, la réverbération, ou le spectre par exemple, mais également des indices dynamiques ou non-acoustiques en fonction des situations. Tous ces indices nous permettent donc au quotidien de percevoir les sources sonores qui nous entourent. Les performances de localisation auditive sont variables d'un individu à l'autre, mais aussi en fonction de la provenance du son.

L'acuité du système visuel est, quant à elle, très élevée dans la zone frontale définie par un cône d'environ 2°. En dehors du champ de vision central, ces performances se dégradent nettement. La perception de la profondeur d'un objet visuel est basée sur l'exploitation de différents indices monoculaires et binoculaires (occlusion, taille relative, perspective linéaire, disparités binoculaires, etc.).

Les systèmes perceptifs auditifs et visuels fournissent des informations multi-sensorielles à notre cerveau. La plupart du temps, la flexibilité de ce dernier permet d'intégrer et d'interpréter ces informations de manière cohérente. Il a été montré dans ce chapitre qu'il existe une fenêtre spatiotemporelle dans laquelle l'intégration audio-visuelle se produit : 100 ms avant et après le point de simultanéité subjective (ce dernier est détecté lorsque le son présente un retard d'environ 30 à 50 ms) et un écart angulaire en azimut inférieur à 3° (dans le cas de stimuli abstraits). Lorsque les stimuli auditifs et visuels sont placés dans cette fenêtre spatiotemporelle, ils sont toujours perçus comme étant spatialement cohérents. Notre système perceptif semble plus tolérant aux écarts temporels et spatiaux dans le cas de stimuli plus réalistes (cas de la parole associée à un personnage par exemple). La perception audio-visuelle résulte donc de processus complexes pouvant impliquer des facteurs sensoriels tels que la cohérence spatiotemporelle des stimuli, mais aussi des facteurs cognitifs comme le lien sémantique, ou la présomption d'unité entre les informations unimodales. Ces facteurs mettent en évidence l'influence des stimuli sur la perception audio-visuelle, tout comme la manière dont ils sont présentés (méthodes et consignes expérimentales).

Toutes ces informations relatives à la perception unimodale et bimodale sont à prendre en compte lors de la conception d'expériences subjectives ou de systèmes de restitution audio-visuelle.

Chapitre 2

Restitution sonore spatialisée et vidéo stéréoscopique 3D

2.1 Introduction

Notre système auditif nous permet d'appréhender notre environnement en trois dimensions. En effet, grâce à nos deux oreilles, nous percevons les objets sonores qui nous entourent, quelles que soient leurs positions dans l'espace. Il est possible de recréer virtuellement cette situation d'écoute grâce aux technologies de spatialisation sonore. Une scène sonore spatialisée offre à l'auditeur la possibilité d'entendre des sons provenant de différents endroits de l'espace, en azimut, en élévation, et en profondeur. Différentes technologies sont capables d'offrir des scènes sonores spatialisées. Ces technologies sont plus ou moins complexes à mettre en œuvre, notamment en fonction du nombre de haut-parleurs utilisés. La partie 2.2 propose une présentation des principales technologies de spatialisation sonore : leur principe de fonctionnement, leurs atouts et leurs défauts.

Les technologies de vidéos stéréoscopiques permettent, quant à elles, de recréer artificiellement le relief d'une scène visuelle. Le principe de la stéréoscopie est apparu au XIX^{e} siècle, et a observé un fort développement à la fin des années 2000, avec la démocratisation du cinéma en relief et l'arrivée des téléviseurs 3D. Les principales technologies vidéo 3D (passive, active, et autostéréoscopique) sont présentées dans la partie 2.3.

Enfin, des exemples d'applications audio-visuelles (télévision, cinéma, réalité virtuelle), associant une présentation visuelle stéréoscopique à des systèmes de spatialisation sonore, sont proposés dans la partie 2.4.

2.2 Technologies de restitution sonore spatialisée

2.2.1 Les technologies multicanales

Les technologies multicanales sont probablement les technologies de spatialisation les plus connues, car les plus répandues. La version la plus ancienne et la plus basique de cette technologie est la stéréophonie (système 2.0) [Blumlein, 1933] qui a été déclinée vers des systèmes plus complexes comme le 5.1, 7.1, 22.2, etc.

Principe

La stéréophonie d'intensité, ainsi que tous les systèmes multicanaux, repose sur le principe du panoramique d'intensité et utilise les indices de latéralisation décrits dans la section 1.2.1 (différences de temps d'arrivée et de niveau sonore). La stéréophonie utilise deux haut-parleurs (droit et gauche), et permet la spatialisation de sources sonores sur la portion d'axe délimitée par ces haut-parleurs grâce au concept de sources virtuelles ou sources fantômes. La source fantôme peut ainsi être placée dans l'intervalle $\theta = [-30°; 30°]$ en appliquant les gains g_R et g_L respectivement aux haut-parleurs droit et gauche. Le signal sonore est alors envoyé aux haut-parleurs d'après la loi des sinus ou la loi des tangentes. Cette dernière est définie par

$$\frac{\tan(\theta)}{\tan(\theta_0)} = \frac{g_L - g_R}{g_L + g_R}, \tag{2.1}$$

où $\theta_0 = 30°$ pour une configuration stéréophonique classique. La restitution stéréophonique optimale a lieu lorsque l'auditeur et les deux haut-parleurs forment un triangle équilatéral. L'auditeur est alors placé au niveau du *sweet spot* comme l'illustre la figure 2.1.

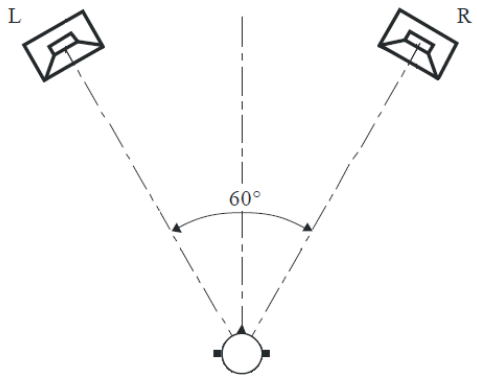


FIGURE 2.1 – *Configuration stéréophonique recommandée par la norme ITU-R BS 775 [ITU 775, 2012].*

Le système 5.1, apparu dans les années 1990, propose une scène sonore plus enveloppante en comparaison de la stéréophonie, grâce à la disposition des 5 haut-parleurs répartis autour de l'auditeur (voir figure 2.2). Le format 5.1 a été le premier format multicanal numérique au cinéma et est utilisé dans la plupart des salles. Son application aux jeux vidéo est apparue dans les années 2000.

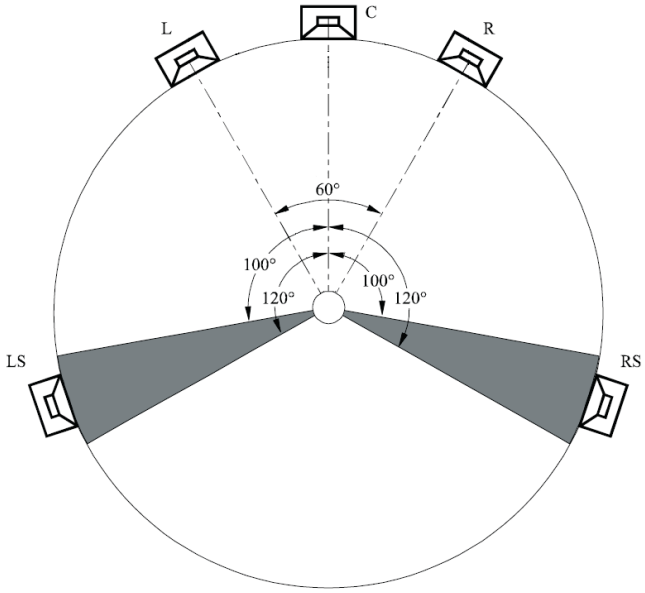


FIGURE 2.2 – *Configuration 5.1 recommandée par la norme ITU-R BS 775 [ITU 775, 2012].*

Il existe également des déclinaisons multicanales impliquant un nombre plus important de haut-parleurs, telles que le 7.1, le 9.1, ou encore le 22.2. Avec l'augmentation du nombre de haut-parleurs, les scènes sonores sont spatialisées en azimut, mais aussi en élévation. Le système 22.2, initialement développé par la NHK [Hamasaki et al., 2004], propose, par exemple, 10 haut-parleurs dans le plan horizontal, auxquels viennent s'ajouter 9 haut-parleurs au-dessus et 3 haut-parleurs en dessous du plan horizontal (voir figure 2.3). Il est à noter que, quelle que soit la configuration multicanale installée, la zone d'écoute est limitée et optimale au *sweet spot*.

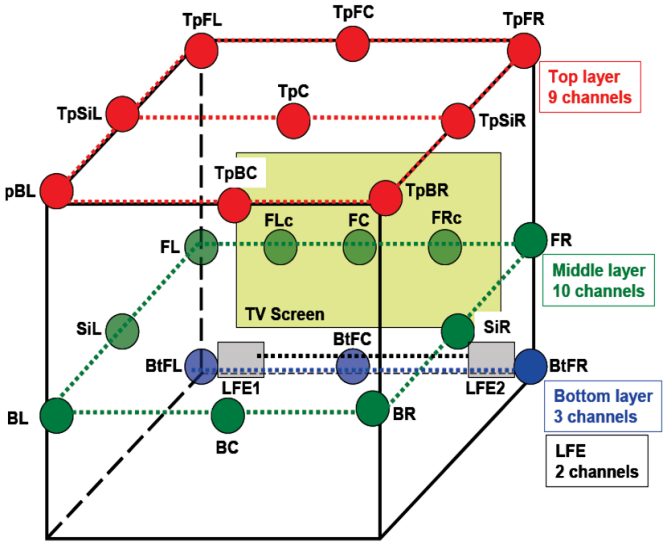


FIGURE 2.3 – *Système 22.2 initialement proposé par la NHK, présenté ici dans sa configuration cubique, d'après Hamasaki [Hamasaki et al., 2004].*

Atouts

La stéréophonie comme le système 5.1 offrent des effets de latéralisation avec un nombre de canaux restreint, ce qui facilite leur mise en œuvre. De plus, ces formats sont très largement répandus dans différents cas d'application (télévision, cinéma, jeux vidéo, etc.). Ils bénéficient à ce titre de systèmes de captation et d'outils de production dédiés.

Défauts

Les principaux défauts des technologies multicanales résident dans la spatialisation sonore qui peut être incomplète et/ou hétérogène suivant l'installation. En effet, la stéréophonie propose une spatialisation uniquement frontale, où les sources virtuelles sont placées entre les haut-parleurs droit et gauche. Les systèmes 5.1 et 7.1 proposent, quant à eux, des effets de spatialisation uniquement horizontale, avec une spatialisation hétérogène en fonction de l'azimut liée à l'écart angulaire entre les haut-parleurs. Les systèmes utilisant un plus grand nombre de haut-parleurs proposent une spatialisation plus fine dans le plan horizontal, ainsi que des informations sonores dans le plan vertical. Pour toutes les installations multicanales, la zone d'écoute est limitée au *sweet spot*, ce qui peut représenter un inconvénient majeur en fonction du type d'application visé. Par ailleurs, les technologies multicanales manquent de flexibilité. En effet, le mixage est figé pour chaque canal dès lors qu'un signal sonore donné est destiné à un haut-parleur placé à une position précise dans l'espace de restitution.

2.2.2 Vector Base Amplitude Panning (VBAP)

Principe

Développé par Ville Pulkki [Pulkki, 1997], le Vector Base Amplitude Panning ou VBAP repose sur le principe de la stéréophonie, appliqué à un triplet de haut-parleurs. La source virtuelle peut alors être placée entre ces trois haut-parleurs en contrôlant le gain de chaque haut-parleur.

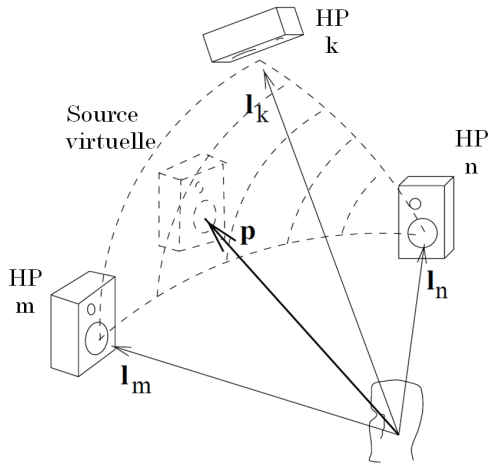


FIGURE 2.4 – *Dispositif VBAP à trois haut-parleurs, d'après Pulkki [Pulkki, 1997].*

Cette technique permet de reconstruire un espace sonore 3D autour de l'auditeur, en faisant appel à plusieurs triplets de haut-parleurs placés sur une sphère. Sur la figure 2.4, si on considère le vecteur

$$l_i = \begin{bmatrix} l_{i1} \\ l_{i2} \\ l_{i3} \end{bmatrix} \tag{2.2}$$

donnant la direction du haut-parleur i, et le vecteur

$$p = \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} \tag{2.3}$$

représentant la direction de la source virtuelle cible, alors il est possible d'écrire la relation

$$p = g_1 l_1 + g_2 l_2 + g_3 l_3, \tag{2.4}$$

où g_1, g_2, et g_3 sont les gains des trois haut-parleurs définis par le vecteur

$$g = L_{123}^T \, p = \begin{bmatrix} l_{11} & l_{21} & l_{31} \\ l_{12} & l_{22} & l_{32} \\ l_{13} & l_{23} & l_{33} \end{bmatrix}^{-1} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} \tag{2.5}$$

à condition que L_{123}^{-1} existe, ce qui est vérifié si L_{123} couvre trois dimensions. Une normalisation des gains peut ensuite être appliquée, en considérant l'amplitude constante (VBAP) ou l'énergie constante (VBIP pour *Vector Base Intensity Panning* [Pernaux et al., 1998, Jot et al., 1999]).

Atouts

Le principal avantage du VBAP réside dans sa simplicité de mise en œuvre.

Défauts

La zone d'écoute est, comme pour les systèmes multicanaux, limitée à un point de l'espace (*sweet spot*). De plus, le VBAP est une technologie expérimentale ne disposant pas de format associé ni de système de captation dédié. Enfin, bien que le VBAP permette de restituer des sources sonores en azimut et en élévation, le rendu de la distance est quant à lui limité avec cette technologie.

2.2.3 La technologie binaurale

Principe

Le principe de la technologie binaurale repose sur une reproduction au niveau du conduit auditif de l'auditeur des informations acoustiques nécessaires à la construction par le système auditif d'une image sonore spatialisée. Le but est de procurer à l'auditeur l'illusion parfaite d'être immergé dans une scène sonore en diffusant une reproduction du champ sonore capté par les oreilles en condition d'écoute naturelle. Si la mise en œuvre de cette technique est réalisée avec soin, l'auditeur perçoit les sons comme provenant de sources nettement éloignées de sa tête, dans des directions bien définies. Cette perception illusoire permet d'approcher au mieux les sensations d'une écoute naturelle.

Encodage binaural

Au quotidien, pour percevoir une scène sonore en trois dimensions, les deux signaux captés au niveau de chaque tympan de l'auditeur suffisent à décrire l'information spatiale du point de vue du système auditif. Les technologies binaurales sont basées sur cette idée : la scène sonore est ainsi représentée par seulement deux signaux qui correspondent aux signaux perçus au niveau des tympans. L'information spatiale est encodée au travers des indices de localisation que sont les différences interaurales de temps (ITD) et d'intensité (ILD) et les indices spectraux (IS). Tous ces paramètres d'encodage dépendent de la position de la source (voir partie 1.2.1).

Les HRTF (Head Related Transfer Function) définissent 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 mis à disposition du système auditif pour localiser les sons, et constituent ainsi le processus fondamental de l'encodage spatial binaural. Les HRTF traduisent l'empreinte morphologique de l'auditeur. Il en résulte que l'encodage binaural est individuel, c'est-à-dire que l'encodage des informations spatiales ne vaut que pour un individu, ce qui est la principale limitation des technologies binaurales.

En pratique, les technologies binaurales se déclinent sous deux formes :

— **encodage naturel** : les signaux binauraux sont acquis par un enregistrement en plaçant une paire de microphones à l'entrée des conduits auditifs d'un individu ou d'un mannequin (tête artificielle). Cette déclinaison trouve son application dans la captation de scènes sonores pour le partage d'ambiance ou le concept de carte postale sonore.

— **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'utilisation des filtres binauraux permet de reproduire au mieux tous les indices de localisation nécessaires, et ainsi d'assurer une illusion satisfaisante.

Décodage binaural

Quel que soit le principe d'encodage binaural utilisé, l'écoute au casque se révèle être la méthode de décodage la plus intuitive. Dans l'étape de décodage, le premier requis est de veiller

à corriger la réponse du casque. Il en résulte la nécessité de calibrer le casque (compenser la fonction de transfert entre le casque et l'entrée des conduits auditifs appelée Head-Phone Transfer Function). Pour un casque donné, les HPTF dépendent principalement de l'individu mais aussi du positionnement du casque [McAnally and Martin, 2002, Kim and Choi, 2005]. Il faut donc, dans l'idéal, appliquer une calibration individuelle. Le choix du casque a aussi son importance [Moller, 1992]. Il est préférable d'utiliser des casques de type ouvert. Ils doivent offrir les conditions de rayonnement en champ libre en termes d'impédance vue par le tympan, comme si l'auditeur ne portait pas de casque.

L'utilisation de haut-parleurs comme système de restitution est également envisageable mais amène le problème des trajets croisés qui entraîne une diaphonie entre les deux oreilles et détruit l'illusion de scène sonore virtuelle. La mise en œuvre du procédé d'annulation des termes croisés est nécessaire [Schroeder and Atal, 1963, Gardner, 1997]. On parle alors de technique transaurale [Atal and Schroeder, 1966, Cooper and Bauck, 1989] ou de stéréo dipôle [Kirkeby et al., 1998b, Kirkeby et al., 1998a] suivant le positionnement des haut-parleurs.

Atouts

La technologie binaurale propose une spatialisation 3D complète et naturelle grâce à uniquement deux canaux. Cet atout positionne sûrement le binaural comme étant la technologie de spatialisation sonore ayant le plus fort potentiel de développement. De plus, la restitution binaurale peut être réalisée sur n'importe quel casque, ce qui la rend compatible avec les terminaux mobiles.

Défauts

Un des freins majeurs au développement du binaural est son caractère individuel. En effet, la qualité de la restitution de la scène sonore en termes de spatialisation, mais aussi de timbre, dépend des caractéristiques physiques et morphologiques de l'auditeur. De plus, l'utilisation de HRTF non-individualisées entraîne généralement un défaut d'externalisation.

2.2.4 Technologie ambisonique

Principe

La technologie de restitution sonore ambisonique est basée sur la décomposition du champ acoustique en harmoniques sphériques centrée sur le point de vue de l'auditeur. Initialement proposée au premier ordre par Gerzon [Gerzon, 1973, Gerzon, 1985], l'approche ambisonique a ensuite été étendue aux ordres supérieurs par Daniel [Daniel, 2000] sous le nom de *Higher Order Ambisonics* (HOA).

La technologie HOA repose sur le développement de l'onde acoustique sur la base des fonctions propres de l'équation des ondes en coordonnées sphériques. Ces fonctions propres combinent des fonctions de Bessel et/ou Hankel sphériques qui décrivent les dépendances radiales, et des harmoniques sphériques qui décrivent les dépendances angulaires de l'onde acoustique

[Bruneau, 1983]. Dans la plupart des cas, le domaine d'écoute est exempt de sources primaires, ce qui permet d'écrire la pression au point $\overrightarrow{r}$ sous la forme simplifiée de série de Fourier-Bessel :

$$p(\overrightarrow{r},\omega) = \sum_{m=0}^{\infty} i^m j_m(kr) \sum_{n=0}^{m} \sum_{\sigma=\pm 1} B_{mn}^{\sigma}(\omega) Y_{mn}^{\sigma}(\phi,\theta), \tag{2.6}$$

où $Y_{mn}^{\sigma}(\theta,\varphi)$ représentent les harmoniques sphériques, $j_m(kr)$ sont les fonctions de Bessel de première espèce d'ordre m, B_{mn}^{σ} les composantes HOA, et $k = \omega/c$ est le nombre d'onde.

Encodage

Le principal avantage de cette technologie réside dans l'encodage du champ sonore qui est réalisé indépendamment du dispositif de restitution. Les microphones de type Soundfield permettent un encodage ambisonique à l'ordre 1 [Craven and Gerzon, 1977] (figure 2.5 (a)). L'encodage est alors réalisé grâce à uniquement quatre composantes (B-format) : W (pression) et X, Y, Z (gradient de pression). Pour les ordres supérieurs, il est possible d'utiliser des sphères microphoniques [Moreau, 2006], comme l'illustre la figure 2.5 (b). Il est à noter que la représentation HOA est une description hiérarchique. En effet, les composantes des premiers ordres suffisent à représenter l'onde acoustique, les ordres supérieurs venant préciser l'information spatiale.

(a) (b)

FIGURE 2.5 – *Microphones ambisoniques ou assimilables : (a) Soundfield (4 capsules pour une restitution ambisonique à l'ordre 1), (b) EigenMike (32 capsules pour une restitution ambisonique à l'ordre 4).*

Décodage

L'étape de décodage consiste à adapter la scène sonore encodée au format HOA à un dispositif particulier de haut-parleurs (nombre et disposition des haut-parleurs). L'espace sonore peut être restitué sur divers dispositifs panoramiques (2D) ou sphériques (3D), avec un nombre de haut-parleurs dépendant de l'ordre M de la décomposition. En 2D, il faut au minimum $(2M+1)$ canaux, et $(M+1)^2$ pour une restitution en 3D. Il est à noter que ce nombre

minimum de haut-parleurs correspond également au nombre optimal, puisque l'ajout de canaux peut entraîner l'apparition d'artefacts audibles [Bertet, 2009]. En augmentant l'ordre de décomposition, la résolution spatiale de la restitution augmente et la zone d'écoute s'étend [Daniel and Moreau, 2004, Bertet et al., 2006, Daniel et al., 2003]. Cette zone d'écoute dépend également de la fréquence : elle est plus étendue pour les basses que pour les hautes fréquences. Prenons l'exemple d'un auditeur placé non pas au centre du dispositif de restitution, mais à une position excentrée. Il est tout à fait possible que cet auditeur soit placé au sein de la zone d'écoute optimale pour des fréquences inférieures à une fréquence de coupure f_c, mais en dehors de cette zone pour les fréquences supérieures à f_c. Une augmentation de l'ordre de décomposition M a pour conséquence l'extension de la zone d'écoute optimale pour toutes les fréquences. Ainsi, pour un auditeur excentré, l'augmentation de l'ordre M peut également entraîner un décalage de la fréquence de coupure f_c vers les hautes fréquences, et donc une augmentation de la bande passante fréquentielle.

La disposition des haut-parleurs peut théoriquement être régulière ou irrégulière. Cependant, une configuration régulière est généralement préférée afin de simplifier la matrice de décodage nécessaire pour adapter les signaux B_{mn}^{σ} au réseau de haut-parleurs.

Atouts

Le découplage entre l'étape d'encodage et de décodage procure à la technologie ambisonique une grande flexibilité. En effet, l'encodage de la scène sonore ne dépendant pas du système de restitution final, il est possible de modifier *a posteriori* la scène sonore encodée pour l'adapter à n'importe quelle configuration de haut-parleurs. De plus, la scène sonore encodée peut être manipulée simplement avec la possibilité d'appliquer une rotation à l'ensemble de la scène lors de la phase de décodage. Un autre point fort de cette technologie est son caractère hiérarchique : les premiers ordres permettent de représenter l'onde acoustique d'un point de vue global, et les ordres supérieurs viennent ensuite préciser l'information spatiale. Cette propriété peut être utile en fonction du contexte d'application et du nombre de haut-parleurs disponibles pour la restitution sonore. Enfin, la technologie ambisonique est depuis peu prise en charge dans le standard international MPEG-H. Le *Draft International Standard* [ISO/IEC DIS 23008-3, 2014] propose en effet que les signaux HOA soient supportés jusqu'à l'ordre 9 en entrée de l'encodeur MPEG-H.

Défauts

Même si elle a tendance à s'étendre avec les ordres supérieurs, la zone d'écoute est cependant restreinte au *sweet spot* pour les premiers ordres (1 et 2). D'autre part, il a été montré que la précision de la spatialisation sonore dépend de l'ordre M de décomposition. Afin de garantir une restitution spatialisée précise en 3D, il est donc nécessaire d'utiliser les ordres supérieurs, et donc un nombre important de canaux ($(M+1)^2$). Notons que des approches hybrides, appelées *Mixed-Order Ambisonics* (MOA), ont été proposées pour limiter le nombre de haut-parleurs en configuration 3D [Marschall and Chang, 2013].

2.2.5 Technologie Wave Field Synthesis

Principe

Le principe de la Wave Field Synthesis (WFS) est initialement proposé par Berkhout [Berkhout, 1988, Berkhout et al., 1993]. La WFS est basée sur le **Principe de Huygens** (illustré sur la figure 2.6) selon lequel un front d'onde généré par une source primaire se comporte comme une distribution de sources secondaires émettant des ondelettes. Dans notre cas, la source primaire correspond à la source sonore virtuelle à restituer, et chacune des ondelettes est générée par un haut-parleur (sources secondaires). Ainsi, la superposition des fronts d'onde des haut-parleurs permet de reconstruire une copie de l'onde acoustique virtuellement émise par la source primaire.

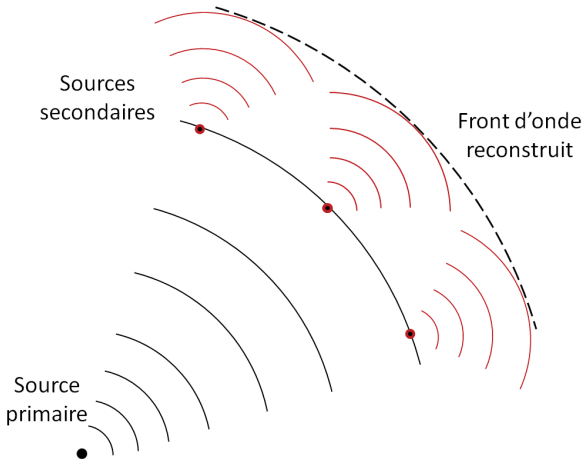


FIGURE 2.6 – *Illustration du Principe de Huygens.*

Mathématiquement, le problème peut être posé sous la forme de l'intégrale de Kirchhoff-Helmholtz. L'espace est décomposé en 2 sous-espaces : un sous-espace Ω_1 contenant les sources acoustiques primaires, et un sous-espace Ω_2 constituant la zone d'écoute. La pression acoustique $p(\overrightarrow{r})$, en tout point $\overrightarrow{r}$ à l'intérieur de Ω_2, est définie grâce à la pression p_0 et au gradient de pression $\overrightarrow{\bigtriangledown}_{p_0}$ sur la surface de séparation entre Ω_1 et Ω_2 (notée $\partial\Omega_0$) par l'équation intégrale

$$p(\overrightarrow{r}) = \iint_{\partial\Omega_0} \left[\overrightarrow{\bigtriangledown}_{p_0}.\overrightarrow{n} - \frac{\overrightarrow{R}}{R}.\overrightarrow{n}(1+jkR)\frac{p_0}{R} \right] \frac{e^{-jkR}}{4\pi R}\, dS_0. \tag{2.7}$$

Dans cette intégrale, le vecteur $\overrightarrow{n}$ représente la normale unitaire à la surface $\partial\Omega_0$ et extérieure au domaine Ω_2, et le vecteur $\overrightarrow{R}$ représente le trajet entre une source secondaire et le point d'écoute. L'intégrale de Kirchhoff-Helmholtz considère donc une distribution continue de sources secondaires, ayant chacune deux composantes : un monopôle (alimenté par le signal à gradient de pression), et un dipôle (alimenté par le signal de pression). Là où le Principe de Huygens impose que les sources secondaires soient positionnées sur le front d'onde de la source primaire, l'intégrale de Kirchhoff-Helmholtz ne fait aucune supposition sur la géométrie de la

surface $\partial\Omega_0$. Cette propriété permet de placer les sources secondaires librement dès lors que les relations de phases entre les sources secondaires sont prises en compte.

Le passage de la théorie à la pratique nécessite cependant de prendre certaines précautions à cause des approximations réalisées en fonction des réseaux de transducteurs considérés :

- La **réduction à une ligne infinie** de sources secondaires entraîne par exemple la **limitation du rendu sonore spatialisé au plan horizontal**. De plus, le champ acoustique observe un **rayonnement à symétrie cylindrique** autour de la ligne de transducteurs. Enfin, l'**atténuation du champ acoustique** est également modifiée et peut être incorrecte en fonction de la position dans la zone d'écoute.
- La **réduction à une ligne finie** de sources secondaires peut également provoquer des artefacts audibles (cas de réseaux linéaires et non circulaires). Un phénomène de **diffraction** est, en effet, occasionné par les extrémités du réseau de transducteurs. Une solution permettant de minimiser ces effets consiste à pondérer les gains des haut-parleurs placés aux extrémités, de manière à diminuer leur contribution [Vogel, 1993]. La deuxième conséquence de la réduction à un segment est la **limitation de la zone d'écoute** dans laquelle il est possible de restituer correctement les sources primaires. Il est à noter que la solution proposée pour limiter les effets de diffraction apporte, pour les mêmes raisons, une limitation de la zone d'écoute [Boone et al., 1995].
- Enfin, quelle que soit la géométrie retenue, il est nécessaire d'effectuer un **échantillonnage** de la ligne de sources secondaires. Cette approximation entraîne l'apparition du phénomène de **repliement spatial** au-delà de la fréquence d'*aliasing* définie par

$$f_{al} = \frac{c}{2\triangle_{tranducteur}}, \tag{2.8}$$

où c est la célérité du son dans l'air, et $\triangle_{tranducteur}$ est la distance entre deux transducteurs (microphones ou haut-parleurs). Un espace de 10 cm entre deux haut-parleurs engendre ainsi des phénomènes de repliement spatial au-delà de $f_{al} = 1700\ Hz$. Start [Start, 1997] a montré que l'incidence du repliement spatial n'est pas critique en termes de localisation sonore, dès lors que la fréquence de repliement spatial est supérieure à 1500 Hz. Start note cependant que, même si la spatialisation sonore est préservée, des artefacts audibles tels que la dégradation du timbre peuvent néanmoins être perçus.

Encodage

En théorie, la formulation de Kirchhoff-Helmholtz impose de capter le champ acoustique émis par des sources primaires en utilisant un réseau de microphones de pression, et à gradient de pression, placés de manière continue sur une surface fermée, délimitant ainsi la zone d'écoute. En pratique, il est possible de disposer un réseau discret de microphones de pression ou à gradient de pression, mais cette méthode reste difficile à mettre en place. De plus, la discrétisation du réseau microphonique (ou échantillonnage spatial) engendre des artefacts au-delà de la fréquence de repliement spatial. Le recours au réseau de microphones peut être évité en utilisant une autre méthode. En effet, il est également possible de capter le champ

acoustique d'une source primaire grâce à un seul microphone, placé à proximité de cette source. Le signal monophonique peut ensuite être envoyé à un réseau virtuel de microphones (ou au réseau de haut-parleurs) en appliquant les relations adaptées de gain et de phase à chaque transducteur.

Décodage

La restitution WFS est assurée par un réseau de haut-parleurs en 2D ou 3D [Corteel et al., 2012, Rohr et al., 2013], théoriquement disposé de la même manière que le réseau microphonique. Comme pour l'étape d'encodage, l'intégrale de Kirchhoff-Helmholtz impose une restitution sur des transducteurs se comportant en monopôle ou dipôle. Dans la pratique, les haut-parleurs (comme les microphones) ne se comportent jamais comme des transducteurs acoustiques parfaits et ont une directivité particulière. De plus, les installations WFS utilisent, par souci de redondance, un seul type de haut-parleurs au lieu de deux. La plupart du temps, les réseaux sont constitués de haut-parleurs montés sur des enceintes closes, qui sont assimilables à des monopôles acoustiques (sur une gamme limitée de fréquence). La géométrie des réseaux de haut-parleurs est souvent circulaire ou linéaire. Comme mentionné précédemment, différents phénomènes peuvent apparaître suivant la géométrie considérée (échantillonnage spatial ou phénomène de diffraction, par exemple).

Atouts

La *Wave Field Synthesis* propose une spatialisation naturelle et complète (2D ou 3D suivant les systèmes de restitution), sur une zone d'écoute étendue. Cette absence de *sweet spot* en fait la technologie de spatialisation privilégiée pour les applications multi-utilisateurs. De plus, la restitution de l'effet de parallaxe acoustique est également intéressante pour créer une scène sonore virtuelle stable, indépendamment de la position d'écoute [Rébillat et al., 2012, Renard, 2000].

Défauts

Le principal défaut de cette technologie est le nombre élevé de haut-parleurs à mettre en place pour assurer une spatialisation complète en 3D. Certains travaux visent à diminuer le nombre de transducteurs, et à les éloigner les uns des autres, tout en maintenant une qualité de restitution sonore optimale [Corteel et al., 2008, Corteel et al., 2012]. Enfin, il est à noter que la WFS bénéfice uniquement d'un système d'encodage virtuel, et non naturel.

2.2.6 Formats de représentation d'une scène sonore

Pour pouvoir passer de l'étape de captation du champ sonore à l'étape de diffusion, il est nécessaire de transporter les informations sonores (spatiales et temporelles). Différents formats de représentation peuvent être utilisés lors de cette étape : les formats *channel-based*, *object-based* ou *scene-based*.

Pour commencer, le format *channel-based* englobe les technologies multicanales, VBAP et la technologie binaurale. Cette représentation impose qu'un signal sonore soit associé à un

haut-parleur du dispositif de restitution, ce qui rend toute manipulation du champ sonore impossible après la phase de mixage. Le dispositif de restitution sonore est également figé.

Ensuite, le format *object-based* également appelé "tout paramétrique" contient autant de signaux sonores que de sources présentes dans la scène à restituer. La WFS peut être considérée comme une technologie *object-based* dès lors que l'encodage est réalisé virtuellement, c'est-à-dire au moyen de prises de son dédiées à chacun des objets sonores. Le dispositif de restitution est flexible et chaque source peut être manipulée de manière indépendante.

Enfin, pour le format *scene-based*, les signaux sonores ne sont associés ni à un haut-parleur du dispositif de restitution ni à une source sonore unique. C'est par exemple le cas des technologies HOA pour lesquelles la scène sonore, constituée de N sources, peut être restituée à l'ordre M sur un réseaux de $(M+1)^2$ haut-parleurs [1]. Le format de représentation *scene-based* permet une manipulation globale de la scène sonore (rotation) mais n'offre pas la possibilité de manipuler chaque source sonore de manière indépendante. Ce format est également appelé *soundfield-based*.

1. Pour une configuration 3D.

2.3 La vidéo stéréoscopique

La vidéo stéréoscopique 3D est une évolution de la vidéo 2D qui consiste à acquérir puis à restituer les informations de profondeur d'une scène visuelle. Pour ce faire, la scène d'origine doit être captée depuis plusieurs angles de vue (au minimum deux). Les images ainsi captées sont ensuite présentées séparément à chacun des yeux lors de la phase de restitution, comme l'illustre la figure 2.7.

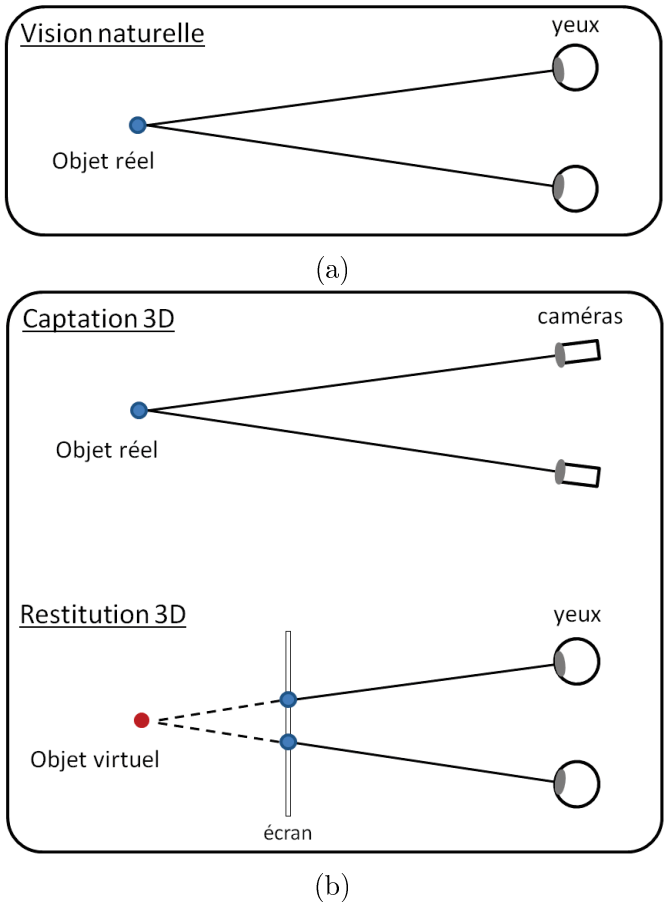


FIGURE 2.7 – *Vision 3D naturelle (a), et étapes de captation et restitution vidéo 3D (b).*

Le décalage entre l'image droite et gauche est appelé disparité stéréoscopique. On distingue trois types de disparité :

- La **disparité nulle**, pour laquelle les deux images ne présentent aucun décalage, permet une visualisation de l'objet virtuel au niveau de l'écran, et donc sans relief.
- La **disparité positive** (ou homonyme) permet de placer l'objet virtuel derrière l'écran.
- Enfin, la **disparité négative** (ou croisée) permet de placer l'objet virtuel entre l'écran et le spectateur. On parle alors d'images en **jaillissement**.

Les différents types de disparité sont illustrés sur la figure 2.8. Plus la disparité entre les images droite et gauche est importante, plus l'objet virtuel est loin de l'écran (derrière l'écran pour les disparités positives, ou en jaillissement pour les disparités négatives).

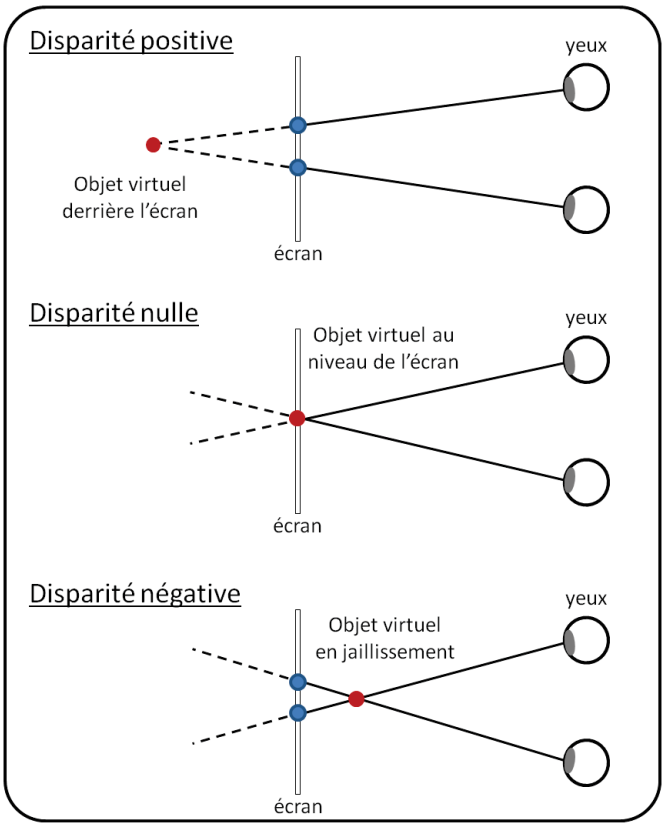


FIGURE 2.8 – *Différents cas possibles de disparités lors d'une restitution vidéo stéréoscopique.*

Afin d'éviter des phénomènes de gêne visuelle lors de la visualisation de contenus 3D, il faut rester dans des gammes de disparités raisonnables. En effet, des phénomènes de divergence peuvent apparaître dès lors que la disparité présentée à l'écran est supérieure à l'écart interoculaire. Ainsi, les objets virtuels doivent être placés dans une gamme de distance appelée boîte scénique, ou *depth budget*, pour garantir un effet 3D optimal.

Le processus de captation / restitution peut être assuré par différentes technologies vidéo 3D. Il est à noter que la perception du relief dépend directement des conditions de captation mais aussi de restitution de la scène 3D. Ces aspects seront abordés après une présentation des principales technologies de restitution 3D.

2.3.1 Les technologies de restitution vidéo 3D

Cette partie a pour but de présenter le principe de fonctionnement des différentes technologies de restitution vidéo 3D classiquement utilisées dans les systèmes destinés au grand public, parmi lesquelles les technologies passives, actives, et autostéréoscopiques [2]. Suivant la technologie employée, la séparation des deux images peut être réalisée par l'intermédiaire de lunettes passives (filtres colorimétriques ou polarisants), de lunettes actives (présentation al-

2. Les technologies basées sur l'holographie ne sont pas abordées dans cette partie puisqu'elles ne sont, pour l'instant, pas disponibles pour le grand public.

ternée des images droites et gauches), ou d'un dispositif placé directement au niveau de l'écran (barrière de parallaxe ou réseau lenticulaire) dans le cas des technologies autostéréoscopiques. La figure 2.9 illustre les différents types de lunettes utilisées dans le cas des technologies 3D passives et actives.

(a) (b) (c)

FIGURE 2.9 – *Lunettes 3D anaglyphes (a), polarisées (b), et actives (c).*

Technologie 3D passive

La 3D passive offre au spectateur la possibilité de voir les images destinées aux deux yeux de manière simultanée. Cette technologie nécessite le port de lunettes spécifiques quelle que soit la méthode mise en œuvre :

— La technologie **anaglyphe** repose sur un filtrage colorimétrique des deux flux vidéo : bleu/rouge, cyan/rouge, ou encore vert/rouge, par exemple. Le spectateur porte des lunettes avec un côté de chaque couleur pour filtrer les images destinées à chaque œil. Cette technique est relativement simple à mettre en place, mais a pour principal inconvénient de perdre les informations de couleur. Des alternatives plus évoluées ont été développées en utilisant des filtres interférométriques par multiplexage de longueurs d'ondes (solution Infitec notamment). Cette solution permet ainsi une reproduction des couleurs plus riche en comparaison de l'anaglyphe, mais occasionne dans tous les cas des déformations colorimétriques entre les deux yeux.

— La technologie 3D à **polarisation** est largement utilisée dans les cinémas, et équipe également certains téléviseurs 3D. Le codage des images droite et gauche est réalisé par l'intermédiaire de deux filtres, polarisés de manière différente, et placés devant les projecteurs dans le cas du cinéma, par exemple. Le spectateur porte des lunettes qui présentent, elles aussi, une polarisation différente entre le côté gauche et le côté droit, afin de bénéficier de la restitution en relief. La polarisation peut être linéaire (horizontale pour l'œil droit et verticale pour l'œil gauche, par exemple), ou circulaire. Cette dernière solution autorise le spectateur à pencher la tête sur le côté tout en bénéficiant de l'effet 3D car la séparation des deux images est maintenue.
La technologie à polarisation offre une 3D de bonne qualité en termes d'espace des couleurs restitué, et de résolution temporelle notamment, avec un dispositif peu onéreux.

En effet, le prix des lunettes en fait une technologie attractive pour les installations destinées au grand public comme le cinéma. Il faut noter cependant que, suivant le type d'installation, la polarisation entraîne une perte de définition de l'image (diminution de la résolution spatiale). En effet, sur les téléviseurs, l'affichage est limité par le nombre de lignes de l'écran (1080 dans le cas d'un téléviseur HD). Pour pouvoir afficher les deux images de manière simultanée, il faut alors présenter chaque image une ligne sur deux, ce qui revient à diviser la résolution verticale par deux. Dans les applications de type cinéma, la définition de l'image n'est pas réduite puisque les deux images sont présentées en pleine définition, grâce à deux vidéo-projecteurs distincts. Par contre, pour ce type d'application, il est nécessaire d'installer un écran de projection métallisé, qui permet de conserver la polarisation différente des deux images projetées. Un autre défaut à noter est que la technologie 3D à polarisation engendre une perte de luminosité importante à cause du filtrage de la lumière par les lunettes.

Technologie 3D active

La 3D active repose sur un affichage séquentiel des images destinées aux deux yeux. Il est ainsi nécessaire de synchroniser le système de diffusion avec les lunettes portées par le spectateur pour qu'à l'instant t, l'image droite soit présentée sur l'écran et que la lunette droite soit active alors que le côté gauche est obturé, et inversement à l'instant $t + \Delta t$. Δt correspond au temps entre deux images successives, et dépend de la fréquence d'affichage du dispositif visuel utilisé [3].

La technologie 3D active est très largement répandue dans les applications destinées au grand public puisque la plupart des téléviseurs 3D en sont équipés. Ce système a pour principal avantage d'offrir des images pleine définition à chaque œil, mais au prix d'une perte de résolution temporelle causée par l'affichage alterné. Si la fréquence d'obturation des lunettes est trop faible, un phénomène désagréable appelé *flicker* peut alors apparaître. De plus, la présence de lumières parasites dans l'espace de visualisation peut également amener des effets de battements à cause de la fréquence d'obturation des lunettes. Ensuite, il a été montré que les technologies actives peuvent engendrer une perte de luminosité plus importante que les technologies passives (lunettes polarisées) [Woods, 2001]. Enfin, les lunettes 3D actives sont plus lourdes que les lunettes passives, puisqu'elles intègrent une batterie et un système de communication pour assurer la synchronisation avec le système de diffusion (TV, projecteur, etc.).

Technologie autostéréoscopique

La technologie autostéréoscopique propose de bénéficier des effets 3D en s'affranchissant du port de lunettes, grâce à une barrière de parallaxe ou un réseau lenticulaire disposé à la surface de l'écran (voir figure 2.10). Les images droite et gauche sont alors diffusées dans différentes directions, ce qui impose au spectateur d'être placé à un endroit précis en face de l'écran. Si le spectateur se décale de cette position optimale, la perception du relief peut alors être perdue ou même inversée (l'œil droit voit l'image gauche et inversement). Quelques téléviseurs permettent

3. $\Delta t = 1/60^{ème} s$ dans le cas d'un affichage cadencé à 60 Hz, par exemple.

une visualisation de contenus 3D suivant plusieurs points de vue, sans cependant permettre un déplacement continu du spectateur. En effet, s'il bouge, le spectateur passe d'un point de vue à l'autre de manière brutale, en passant dans des zones où le relief est inversé. Il faut noter que ces systèmes de restitution à N points de vue nécessitent également une captation suivant N points de vue. Cette contrainte a considérablement limité le développement des technologies autostéréoscopiques.

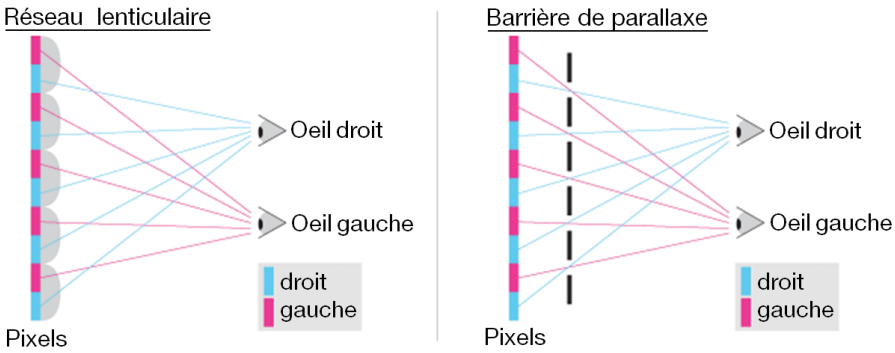


FIGURE 2.10 – *Principe de la technologie autostéréoscopique utilisant une barrière de parallaxe ou un réseau lenticulaire disposé à la surface d'un écran.*

2.3.2 Perception du relief

La partie 1.3.1 de ce document présente les indices visuels monoculaires et binoculaires qui sont interprétés par notre cerveau, dans une situation de visualisation naturelle, de manière à fournir une représentation de l'espace en profondeur. Le but des technologies 3D est de proposer au système perceptif du spectateur une sensation de profondeur se rapprochant d'une situation de visualisation réelle, grâce notamment aux disparités binoculaires. Il a été montré que cet indice est particulièrement important pour la compréhension de la profondeur d'une scène visuelle, dès lors que les objets d'intérêt sont situés à moins de 3 mètres du spectateur (voir figure 1.7). La restitution des disparités binoculaires doit donc être réalisée avec précision, car cette zone de l'espace est généralement exploitée dans les applications vidéo 3D.

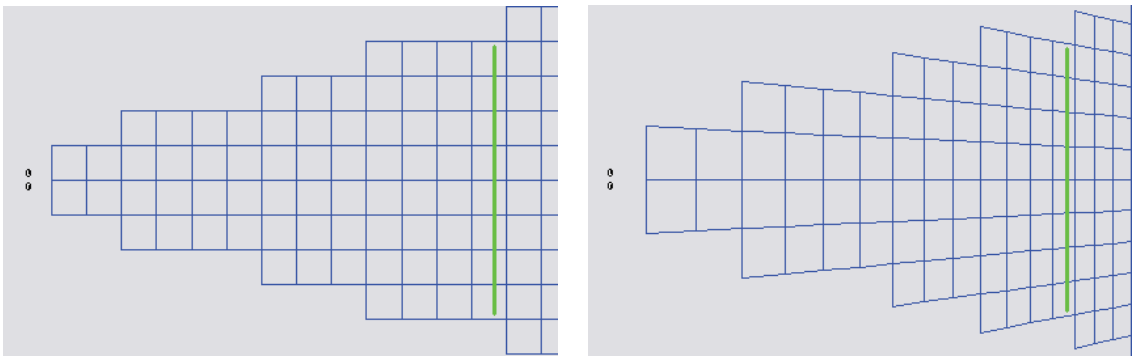
La perception du relief d'images stéréoscopiques est d'autant plus complexe à maîtriser qu'elle dépend à la fois des paramètres de captation de la scène visuelle d'origine et des conditions de visualisation.

Différents paramètres de prise de vue ont un impact sur la perception finale du relief d'une scène visuelle :

— L'écart entre les caméras peut engendrer une restitution du relief non-linéaire, comme le montre la figure 2.11. La plupart du temps, l'écart inter-caméras est égal à la distance interoculaire moyenne, c'est-à-dire 65 mm.
— Le plan de convergence des caméras et la longueur de focale des capteurs peuvent également modifier le relief restitué.
— Pour garantir une visualisation 3D confortable, les éléments de la scène visuelle à capter

doivent être placés dans une gamme de distance définie par la **boîte scénique** (zone de confort). Ses limites sont déterminées en calculant la disparité maximale à ne pas dépasser pour l'objet le plus proche du spectateur (disparité négative) et l'objet le plus loin (disparité positive). En effet, plus les disparités stéréoscopiques sont importantes, plus la différence entre le plan de convergence et le plan d'accommodation est importante. Notre système perceptif est capable de traiter ces conflits dans une certaine mesure. Il est possible de quantifier le conflit vergence-accommodation grâce au critère DoF (*Depth of Focus*). La zone de confort est généralement définie pour un critère DoF égal à ±0.2 dioptrie [Yano et al., 2004, Chen, 2012].

— Enfin, la présence d'asymétries entre les deux caméras lors de la phase de captation peut entraîner de l'inconfort ou de la fatigue lors de la visualisation, et ainsi dégrader la perception du relief. Ces asymétries peuvent être temporelles, optiques (différence de focale entre les capteurs), colorimétriques, lumineuses, ou géométriques (décalage vertical, horizontal, ou encore rotation des caméras) [Balter et al., 2008, Chen, 2012].



(a) (b)

FIGURE 2.11 – *Exemple de distorsion stéréoscopique engendrée par un écart inter-caméras de 40 mm (b) au lieu de 65 mm (a). Visualisation générée avec le logiciel StereoCalculator (outil interne à Orange Labs). La ligne verte verticale représente la position de l'écran.*

Les conditions de visualisation de contenus 3D ont également un impact sur la perception du relief par le spectateur. Parmi ces paramètres, la distance de visualisation est directement liée à la perception de la profondeur. Plus le spectateur est distant de l'écran, plus le relief perçu sera compressé. La taille et la résolution de l'écran peuvent également amener des effets de relief différents.

Conditions de conformité totale, relative, ou partielle

Il a été montré que, suivant le jeu de paramètres de captation et de restitution utilisé, la perception du relief et des dimensions des objets est différente. Il est possible de reproduire le relief et les dimensions d'objets en conformité totale avec la réalité, grâce à un jeu spécifique de paramètres. Il faut pour cela que la distance inter-caméras soit égale à l'écart interoculaire, que la distance de convergence des caméras soit égale à la distance de visualisation, et que l'angle de prise de vue soit égal au champ visuel lors de la restitution. Dès lors que ces critères

ne sont pas respectés, le relief est restitué avec des non-linéarités : on parle alors de conditions de conformité relative ou partielle (voir figure 2.12).

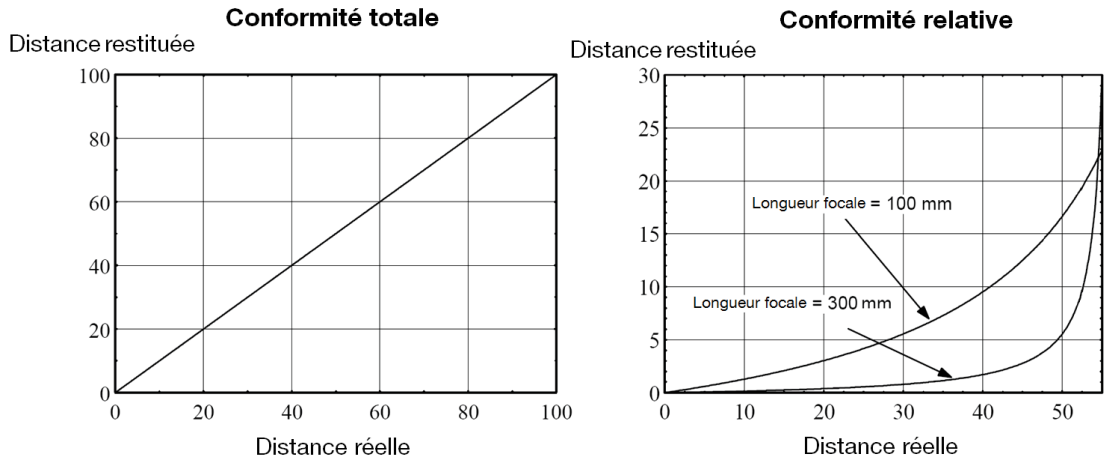


FIGURE 2.12 – *Illustration de la conformité totale et de la conformité relative d'après [Balter et al., 2008]. La relation entre la distance restituée et la distance réelle est linéaire dans le cas de la conformité totale.*

2.4 Technologies de spatialisation sonore et installations audio-visuelles grand public

Différentes technologies de spatialisation sonore sont intégrées dans des systèmes audio-visuels destinés au grand public et couvrent de nombreux domaines d'application.

Cinéma

Depuis les années 1990, le format de restitution sonore multicanal 5.1 est généralement utilisé dans les salles de cinéma. Depuis 2012, certains cinémas s'équipent de la technologie Dolby Atmos qui propose une approche hybride entre une restitution orientée *channel*, et une restitution orientée *object*. Ce système permet ainsi de reproduire simultanément jusqu'à 128 objets sonores dans le plan horizontal mais aussi vertical.
Certains cinémas disposent d'un système de restitution *Wave Field Synthesis*. Le *Grand Cinema Digiplex* à Bucarest (Roumanie), le *Chinese 6 Theatres* à Hollywood (États-Unis), ou encore le *CGV Cinema* à Shenyang (Chine) disposent par exemple de cette technologie.

Télévision et *home cinéma*

Le petit écran bénéficie aussi de systèmes de reproduction sonore basés sur le format 5.1 : on parle de systèmes *home cinema*. La démocratisation de supports tels que le DVD ou le Blu-ray a accentué cette prédominance du 5.1 dans les installations domestiques.

Le format multicanal 22.2 a été développé par la NHK pour accompagner la télévision Ultra Haute Définition (UHDTV 8K ou *super high vision*) qui propose une image 16 fois plus grande

qu'une image haute définition, soit une résolution de 7680 par 4320 pixels [ITU 2020, 2014]. La diffusion de contenus 8K ne sera pas opérationnelle avant 2020, d'après le consortium européen DVB.

Certaines barres de son utilisent désormais le principe de fonctionnement de la technologie Wave Field Synthesis. Depuis 2014, Sonic Emotion intègre ce type de traitement dans certains équipements de la marque Samsung par exemple.

Jeux vidéo

Le format 5.1 est apparu dans les années 2000 dans les jeux vidéo. La plupart des plateformes de jeu proposent des jeux vidéo offrant plusieurs mixages audio : un stéréo, un 5.1, et, plus rarement, un mixage 7.1.

Développé par Crytek, le jeu *Crysis 3* propose en 2013 un niveau dont le mixage audio a été réalisé avec la technologie Proximity de Iosono [Iosono, 2014]. Cette technologie, basée sur la Wave Field Synthesis, offre une restitution sonore au casque, ou bien grâce à deux barres de son placées de part et d'autre de la pièce. Le jeu *Crysis 3* s'est vendu à 260 000 exemplaires aux États-Unis, dans les 12 jours qui ont suivi son lancement [Gamespot, 2014].

La technologie binaurale fait depuis peu son apparition dans le domaine du jeu par l'intermédiaire des applications mobiles notamment. Sorti en 2011, le jeu *Papa Sangre* propose par exemple de naviguer dans un environnement virtuel en se basant sur les indices sonores distillés grâce à la technologie binaurale. Une deuxième version de ce jeu est sortie en 2013 sous le nom de *Papa Sangre II*.

Évènementiel

Certaines salles de concert utilisent la Wave Field Synthesis comme système de restitution sonore. On peut trouver de telles installations en France, par exemple, à l'Institut du Monde Arabe, ou encore à l'Espace de projection de l'IRCAM. Ce dernier exemple a la spécificité d'associer un système Wave Field Synthesis constitué de 264 haut-parleurs (dans le plan horizontal) à un système de restitution HOA dédié à la reproduction sonore de sources en élévation (dôme de 75 haut-parleurs) [Noisternig et al., 2013].

La diffusion d'évènements culturels est un autre exemple d'application audio-visuelle permettant d'associer du son spatialisé à un contenu vidéo. En 2011, par exemple, une captation de "L'enlèvement au Sérail" de Mozart a été réalisée à l'Opéra de Rennes [Nicol et al., 2013]. L'évènement a été diffusé en direct sur iPad avec une restitution sonore binaurale. Deux ans plus tard, l'Opéra de Rennes a organisé une nouvelle captation audio-visuelle ("La Traviata" de Verdi). Cette expérience a été l'occasion d'utiliser une restitution binaurale en association avec une vidéo à 360 ° présentée sur iPad.

2.5 Conclusion

L'objet de ce chapitre était de présenter les différentes technologies de spatialisation sonore et de rendu stéréoscopique 3D.

Il existe de nombreuses technologies de restitution sonore parmi lesquelles les technologies multicanales, le VBAP, le binaural, l'ambisonique, et la Wave Field Synthesis. Ces technologies permettent la restitution de champs sonores dans le plan horizontal et/ou vertical et peuvent, suivant les cas, impliquer entre deux et plusieurs centaines de transducteurs.

Les technologies de vidéos stéréoscopiques fournissent au spectateur une impression de relief visuel. La démocratisation de ces technologies a été portée par la création de nombreux contenus 3D conjointement à la production de téléviseurs 3D depuis la fin des années 2000. Contrairement aux technologies autostéréoscopiques, les technologies de vidéos 3D passives et actives nécessitent le port de lunettes spécifiques. Bien qu'il existe des différences importantes entre toutes ces technologies, la perception du relief dépend principalement des conditions de captation et de restitution des contenus vidéo.

Nous avons montré que la spatialisation sonore trouve sa place dans de nombreuses applications audio-visuelles destinées au grand public (cinéma, télévision, jeux vidéo, etc.). Dans la plupart des cas, la chaîne de diffusion sonore est néanmoins limitée au format 5.1. La tendance semble indiquer que les formats *object-based* vont prendre de plus en plus d'importance grâce à leur gestion dans les codeurs audio (MPEG-H, par exemple), mais aussi grâce à la création d'outils de mixage et de post-production dédiés. Même si les instances de standardisation telles que l'ITU et l'EBU travaillent sur la définition des futurs formats audio, le système de restitution sonore 5.1 reste le format recommandé pour la visualisation de vidéos 3D [EBU R-135, 2012].

Chapitre 3

Qualité d'expérience audio-visuelle 3D

3.1 Introduction

La qualité d'un signal ou d'un service peut être évaluée suivant deux approches principales.

La première regroupe les méthodes d'évaluation objective qui utilisent les mesures physiques des signaux audio ou vidéo. Ces méthodes permettent ainsi de caractériser objectivement la qualité d'un signal ou d'un service en se basant sur des critères tels que le débit transmis ou la perte de paquets audio et/ou vidéo, par exemple. Cette caractérisation permet de déterminer la qualité de service (ou QoS pour *Quality of Service*) qui est définie par l'Union Internationale des Télécommunications (UIT) comme étant "l'ensemble des caractéristiques d'un service de télécommunication qui lui permettent de satisfaire aux besoins explicites et aux besoins implicites de l'utilisateur du service." [ITU E800, 2008]. Les paramètres physiques des signaux multimédias et la qualité de service qui en résulte peuvent être utilisés pour estimer la qualité de service qui sera perçue par les utilisateurs grâce à des modèles de prédiction.

La seconde approche regroupe les méthodes d'évaluation subjective et se focalise sur la qualité des signaux audio ou vidéo telle que perçue du point de vue de l'utilisateur final. Ces méthodes d'évaluation nécessitent la mise en place de tests subjectifs et permettent d'approcher la qualité d'expérience (ou QoE pour *Quality of Experience*) associée à la présentation d'un contenu à travers la qualité média perçue. D'après l'UIT, la qualité d'expérience qualifie "l'acceptabilité globale d'une application ou d'un service tel que subjectivement perçue par l'utilisateur final" [ITU P10, 2008].

Dans ce chapitre, nous allons principalement nous intéresser à la seconde approche, en commençant par une description des méthodes d'évaluation subjective couramment utilisées pour évaluer des contenus audio, vidéo, ou audio-visuels (section 3.2). Les limitations de ces méthodes pour évaluer des séquences audio-visuelles 3D sont également exposées. Ensuite, nous présentons une expérience subjective, réalisée dans le cadre de ces travaux, qui a pour but d'évaluer l'expérience audio-visuelle 3D telle qu'elle peut être présentée actuellement, au cinéma ou à la télévision (section 3.3).

3.2 Méthodes d'évaluation subjective de la qualité audio-visuelle

3.2.1 Évaluation de la qualité audio

Il semble indispensable que la communauté scientifique utilise des méthodes normalisées pour la comparaison des résultats obtenus lors de tests d'écoute subjectifs. Dans ce but, différentes recommandations existent et décrivent la méthodologie à suivre lors de la conception de tests d'écoute. Les méthodes d'évaluation normalisées par l'Union Internationale des Télécommunications telles que MUSHRA [ITU 1534, 2014], UIT-R BS 1116 [ITU 1116, 2014] ou encore UIT-R BS 1284 [ITU 1284, 2003] sont parmi les plus utilisées.

La recommandation UIT-R BS 1116

De nombreux tests d'écoute sont réalisés conformément à la norme UIT-R BS 1116 intitulée : "Méthodes d'évaluation subjective des dégradations faibles dans les systèmes audio y compris les systèmes sonores multivoies" [ITU 1116, 2014]. Cette recommandation est destinée à l'évaluation de la qualité des systèmes et des contenus audio (type de codage, modèle d'enceinte ou encore algorithme de synthèse vocale par exemple) qui introduisent des dégradations infimes et difficiles à déceler. Pour rassembler des informations fiables dans le cas de faibles dégradations, il faut recourir à des méthodes expérimentales strictes, mais aussi à des analyses statistiques appropriées.

Il convient de sélectionner des participants "experts", c'est-à-dire des auditeurs qui ont acquis une expérience dans l'évaluation de contenus audio faiblement dégradés et des capacités propres à la détection de ces défauts. Cette qualification "d'expert" vient en opposition aux auditeurs dits "non experts", ou "naïfs", qui participeront à des tests où les dégradations sont plus franches et ainsi plus faciles à déceler. Dans les multiples expériences qui utilisent la recommandation UIT-R BS 1116, un panel composé d'environ vingt testeurs permet généralement d'obtenir des résultats pertinents.

La méthode d'évaluation UIT-R BS 1116 est une méthode utilisant "un triple stimulus avec une référence cachée". Cette méthode, reconnue comme étant particulièrement sensible et stable, permet de détecter avec précision les dégradations en suivant la démarche détaillée ci-après. Trois stimuli, notés "A", "B" et "C", sont présentés au testeur. Le signal non dégradé (référence) est connu et toujours présenté comme stimulus "A". L'un des stimuli, "B" ou "C", est la référence appelée référence cachée. L'auditeur doit reconnaître cette référence cachée entre "B" et "C" et ensuite noter la version dégradée de l'extrait sur l'échelle de notation.

L'échelle de notation utilisée est continue et représente les niveaux de dégradation perçus allant de "très gênant" à "imperceptible" (voir figure 3.1).

Chaque extrait peut être répété autant de fois que nécessaire avant l'attribution de la note. Lorsque le participant a fini de noter l'extrait, il passe à l'extrait suivant. Le rythme de la procédure de test est donc déterminé par le testeur lui-même. La norme recommande cependant que les extraits sonores présentés à l'auditeur aient une durée comprise entre 10 et 25 secondes, dans le but de limiter le temps total du test (moins de 30 minutes dans l'idéal) et ainsi d'éviter un biais causé par la fatigue du participant.

La méthode d'évaluation décrite ici stipule également avec précision les conditions d'écoute à respecter. Ces conditions d'écoute englobent différents aspects tels que :

- **les propriétés géométriques du local de test :** la surface au sol, la forme et les proportions du local,
- **les propriétés acoustiques du local de test** comme le temps de réverbération,
- **les caractéristiques du champ sonore de référence :** on trouve des indications concernant le son direct, le son réfléchi, mais aussi le bruit de fond,
- **le niveau d'écoute**,
- **la disposition du système d'écoute pour les reproductions monophoniques, stéréophoniques ou multivoies :** position et orientation des haut-parleurs, etc.

Grâce à toutes ces informations, le champ sonore au point d'écoute est rigoureusement déterminé et rend la procédure totalement reproductible.

La méthode MUSHRA

Une autre méthode d'évaluation subjective de la qualité sonore est la recommandation UIT-R BS 1534, plus connue sous le nom de la méthode MUSHRA (*MUlti Stimulus test with Hidden Reference and Anchor*) [ITU 1534, 2014]. La méthode MUSHRA est utilisée pour évaluer les contenus de qualité audio intermédiaire. Il existe plusieurs points communs entre cette méthode et la méthode UIT-R BS 1116, notamment en ce qui concerne la sélection des participants ou encore les conditions d'écoute à respecter lors du test.

C'est principalement dans le protocole de test que la méthode MUSHRA se distingue. Il s'agit d'une méthode à "multi-stimuli avec référence et ancrages cachés". Le premier signal d'ancrage caché correspond au signal sonore original ayant subi une dégradation importante (filtrage passe-bas à $3,5\ kHz$). Cette version est appelée "ancrage de basse qualité". La récente révision de la norme UIT-R BS 1534 mentionne la nécessité d'intégrer un deuxième signal d'ancrage parmi les stimuli à évaluer : "l'ancrage de qualité moyenne". Cet ancrage correspond au signal de référence filtré passe-bas à $7\ kHz$.

Lors d'une phase de notation, toutes les versions disponibles du signal sonore sont présentées au participant de manière simultanée. Par exemple, si un test porte sur la qualité de cinq codages audio, le sujet évalue 9 signaux différents (le signal de référence, cinq signaux dégradés par les codages, la référence cachée et les deux signaux d'ancrage cachés). L'auditeur doit évaluer la qualité de tous les signaux en fonction du signal de référence. La notation est effectuée sur une échelle de qualité continue allant de 0 à 100 (CQS, *Continuous Quality Scale*) présentant cinq intervalles : "mauvais, médiocre, moyen, bon, excellent" (voir figure 3.1). Le critère d'évaluation conseillé est celui de la "qualité audio de base" qui est défini par l'UIT comme incluant "tous les aspects de la qualité du son en cours d'évaluation. Elle inclut entre autres le timbre, la transparence, l'image stéréophonique, la présentation spatiale, les caractéristiques de réverbération, les échos, la distorsion harmonique, (...)" [ITU 1284, 2003]. Cependant, suivant le système de diffusion sonore utilisé, d'autres critères d'évaluation spécifiques peuvent être considérés en complément de la "qualité audio de base" (qualité de l'image stéréophonique, qualité de l'image sonore frontale ou encore qualité de l'impression ambiophonique pour

caractériser la notion d'espace sonore).

La méthode MUSHRA a pour principal avantage de présenter l'ensemble des stimuli simultanément. Ceci permet au participant de faire directement toutes les comparaisons de son choix tout en réduisant la durée du test. La cohérence des résultats s'en trouve accrue, ce qui conduit à des intervalles de confiance plus faibles [ITU 1534, 2014].

La recommandation UIT-R BS 1284

La recommandation UIT-R BS 1284 s'intitule : "Méthodes générales pour l'évaluation subjective de la qualité sonore" [ITU 1284, 2003]. Cette méthode de test propose trois échelles différentes : l'échelle de dégradation à cinq niveaux de la méthode UIT-R BS 1116, l'échelle de qualité à cinq niveaux de la méthode MUSHRA et enfin une échelle de comparaison à sept niveaux. Toutes les échelles sont continues, et la note est définie à la décimale près. La figure 3.1 illustre les différentes échelles proposées dans la recommandation UIT-R BS 1284.

Qualité		Dégradation	
5	Excellent	5	Imperceptible
4	Bon	4	Perceptible, mais non gênant
3	Moyen	3	Légèrement gênant
2	Médiocre	2	Gênant
1	Mauvais	1	Très gênant

Comparaison	
3	Bien meilleure
2	Meilleure
1	Légèrement meilleure
0	Identique
–1	Légèrement plus mauvaise
–2	Plus mauvaise
–3	Bien plus mauvaise

FIGURE 3.1 – *Échelles de notation de qualité, de dégradation et de comparaison, telles que définies dans la norme UIT-R BS 1284 [ITU 1284, 2003].*

D'après cette norme, le test peut se présenter sous différentes formes : une présentation du signal unique, une comparaison par paire (avec ou sans référence parmi les deux signaux), ou encore une comparaison multiple (avec ou sans référence).

La particularité de la méthode UIT-R BS 1284 réside dans l'utilisation de l'échelle de comparaison, mais aussi dans la prise en compte d'attributs pour décrire la qualité sonore perçue. On peut citer quelques attributs pouvant être utilisés tels que l'impression spatiale, la balance ou la couleur du son, par exemple. Cependant, dans la plupart des tests d'écoute utilisant la méthode UIT-R BS 1284, l'auditeur est invité à évaluer la "qualité audio de base" uniquement.

Les différentes méthodes d'évaluation subjective décrites ci-dessus semblent cependant limitées lorsqu'il s'agit d'évaluer la qualité de sons spatialisés grâce à des technologies autres que multicanales. Cette limitation est discutée dans la section 3.2.4 de ce document.

3.2.2 Évaluation de la qualité vidéo

L'Union Internationale des Télécommunications recommande l'utilisation de certaines normes pour évaluer la qualité de vidéos 2D [ITU 500, 2012, ITU 1788, 2007], et de vidéos 3D [ITU 2021, 2012].

Méthodologies d'évaluation de la qualité de vidéos 2D

Les normes UIT-R BT 500 et UIT-R BT 1788 s'intitulent respectivement "Méthodologie d'évaluation subjective de la qualité des images de télévision" [ITU 500, 2012] et "Méthode d'évaluation subjective de la qualité vidéo dans les applications multimédias" [ITU 1788, 2007]. Elles ont pour but de fournir un cadre commun aux études subjectives portant sur la qualité d'images 2D présentées à la télévision ou pour des applications multimédias. Les conditions de présentation y sont définies, ainsi que les protocoles d'évaluation subjective.

Les conditions de présentation des séquences vidéo sont stipulées dans la recommandation UIT-R BT 500 et portent notamment sur :

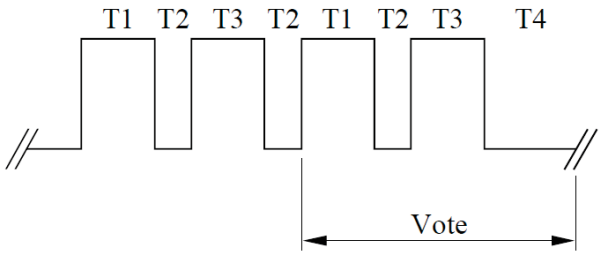
- **les conditions générales de visualisation** telles que la luminosité de la salle et de l'écran, la résolution et le contraste de l'écran, la distance de visualisation, la chromaticité de l'arrière plan, etc.,
- **la sélection des stimuli** et plus particulièrement le nombre et le type de séquences vidéo,
- **le nombre d'observateurs**, qui est fixé au minimum à 15 sujets "non spécialistes",
- **la durée des sessions de test**.

Les protocoles d'évaluation proposés dans cette norme sont nombreux et peuvent être regroupés en trois groupes lorsqu'il s'agit d'évaluer des séquences vidéo courtes (environ 10 secondes) : les méthodes à double stimulus (DSCQS pour *Double-Stimulus-Continuous-Quality-Scale*, et DSIS pour *Double-Stimulus-Impairment-Scale*), les méthodes à un seul stimulus et les méthodes de comparaison. Les échelles de notation utilisées dans ces différentes méthodes sont identiques à celles présentées sur la figure 3.1.

- **Les méthodes à double stimulus (DSCQS et DSIS)**. La méthode DSCQS consiste à présenter une paire d'images (l'une après l'autre) deux fois de suite. L'une des séquences est la référence, l'autre a été traitée par le système à évaluer. Suite à cette présentation répétée, l'observateur doit évaluer la qualité vidéo globale des deux images sur une échelle de qualité (voir figure 3.1). L'ordre de présentation décrit ici est illustré sur la figure 3.2. La méthode DSIS est comparable à la méthode DSCQS à ceci près que la notation est effectuée sur une échelle de dégradation (voir figure 3.1).
- **Les méthodes à un seul stimulus (SS)**. Ces méthodes de présentation sont plus simples puisque l'observateur évalue la qualité vidéo globale de chaque séquence vidéo

à l'essai, sans avoir de référence de haute qualité associée.

— **Les méthodes de comparaison de stimulus (SC)**. Les méthodes de comparaison impliquent la présentation d'une série de paires d'images. Toutes les combinaisons de séquences sont évaluées par le testeur à l'aide de l'échelle de comparaison présentée sur la figure 3.1.



Phases de la présentation:

T1 =	10 s	Image de référence
T2 =	3 s	Gris moyen produit par un niveau vidéo d'environ 200 mV
T3 =	10 s	Condition à l'essai
T4 =	5-11 s	Gris moyen

FIGURE 3.2 – *Structure de la présentation des séquences de test pour les méthodes à double stimulus, d'après [ITU 500, 2012].*

La norme UIT-R BT 500 préconise d'utiliser **les méthodes d'évaluation continue avec stimulus unique (SSCQE) ou à double stimulus simultané (SDSCE)** dans le cas particulier de séquences longues (comprises entre 3 et 5 minutes). Dans ces méthodes, l'observateur évalue de manière continue les variations de qualité d'une vidéo, en présence (SDSCE) ou non (SSCQE) de la référence.

La recommandation UIT-R BT 1788 ("Méthode d'évaluation subjective de la qualité vidéo dans les applications multimédias") propose d'utiliser les méthodes de la norme UIT-R BT 500 ou une méthode additionnelle appelée SAMVIQ (*Subjective Assessment Methodology for Video Quality*). Cette méthode est au domaine de la vidéo ce que la méthode MUSHRA est au domaine de l'audio. En effet, la méthode SAMVIQ propose une présentation multi-stimuli avec une référence explicite et une référence cachée parmi les conditions à l'essai. Les observateurs ont donc accès à plusieurs versions d'une même séquence et doivent évaluer toutes ces versions sur une échelle de qualité continue avant d'accéder au contenu de la séquence suivante.

La recommandation UIT-R BT 1788 préconise des conditions de visualisation comparables à celles mentionnées dans la norme UIT-R BT 500. Des adaptations sont toutefois proposées pour tenir compte de son contexte d'application spécifique, orienté vers le multimédia (distance de visualisation variable en fonction de l'application visée par exemple).

Méthodologies d'évaluation de la qualité de vidéos 3D

Les méthodologies d'évaluation abordées précédemment doivent être en partie repensées dans le cas de la vidéo 3D. L'influence des conditions de visualisation sur la perception du relief a, par exemple, été évoquée dans la section 2.3.2. La recommandation UIT-R BT 2021 intitulée "Méthode d'évaluation subjective de système TV3D stéréoscopiques" est dédiée à l'évaluation subjective de contenus vidéo 3D [ITU 2021, 2012]. Cette norme reprend les principaux éléments de la norme UIT-R BT 500 détaillés précédemment (critères d'évaluation, conditions de visualisation, protocoles d'évaluation, etc.), auxquels des adaptations sont apportées pour s'adapter aux contenus stéréoscopiques.

Deux **critères d'évaluation** spécifiques à l'évaluation des images stéréoscopiques sont, par exemple, ajoutés au critère de la qualité d'image utilisé pour les images 2D : la qualité de la profondeur et le confort visuel. La prise en compte de ces critères permet à la fois de qualifier la qualité d'une vidéo 3D d'un point de vue global et d'évaluer les aspects liés à l'ajout de profondeur visuelle. Cette nouvelle dimension visuelle peut en effet entraîner des phénomènes de gêne visuelle ou de fatigue (notamment dans le cas de séquences longues) qui doivent pouvoir être pris en compte lors d'une phase d'évaluation subjective.

Différentes études se sont intéressées aux **conditions générales de visualisation** dans le cas d'images 3D. Chen et al. [Chen et al., 2010] ont par exemple recommandé de prendre en compte la perte de luminosité induite par le port de lunettes 3D. Ces pertes atteignent jusqu'à 80 % pour des lunettes actives et 60 % pour des lunettes polarisées. Les auteurs mentionnent également un risque d'inconfort visuel lié à l'utilisation de lumières à néon lors des tests subjectifs. En effet, l'utilisation de ce type d'éclairage pourrait provoquer des phénomènes de battement avec l'utilisation de la technologie 3D active. D'autres facteurs à risques sont évoqués par ces mêmes auteurs quant au choix de la position et de la distance de visualisation. Malgré ces recommandations, la norme UIT-R BT 2021 stipule que les conditions générales de visualisation de contenus 3D doivent rester identiques aux conditions de visualisation de contenus 2D (définies dans l'UIT-R BT 2022 [ITU 2022, 2012]).

Certains facteurs propres à l'utilisation de technologies 3D sont néanmoins intégrés dans la norme UIT-R BT 2021. Les limites de disparité à ne pas dépasser sont par exemple mentionnées afin de réduire le conflit entre la distance de convergence et la distance d'accommodation. La norme indique que les objets visuels doivent être placés dans une boîte scénique qui respecte le critère DoF (*Depth of Focus*) de ± 0.2 à ± 0.3 dioptrie (voir section 2.3.2). Ensuite, le nombre minimum d'observateurs pour la vidéo 3D est fixé à 30 participants, soit deux fois plus que pour la vidéo 2D. Cette recommandation tient compte de la variabilité inter-individuelle importante dans le cas d'images 3D [Ukai and Howarth, 2008]. Un autre exemple de recommandation propre à l'évaluation de la vidéo 3D concerne la phase de sélection des participants. Les observateurs doivent réaliser un test de vision stéréoscopique (test de Randot, par exemple) en plus des tests d'acuité visuelle et de perception des couleurs utilisés dans les tests subjectifs portant sur la qualité de vidéos 2D.

Les **protocoles d'évaluation** sont quant à eux les mêmes que pour l'évaluation de la

qualité de vidéos 2D. La méthode à double stimulus DSCQS, la méthode de comparaison, la méthode à un seul stimulus et la méthode d'évaluation continue à un seul stimulus SSCQE sont considérées comme pertinentes pour évaluer la qualité de l'image, la qualité de la profondeur et le confort visuel. Ces protocoles d'évaluation impliquent l'utilisation des échelles de notation de qualité et de comparaison. D'après l'UIT, les protocoles utilisant l'échelle de dégradation ne sont pas pertinents dans ce contexte. Seule l'échelle de notation de qualité est modifiée pour l'évaluation du critère lié au confort visuel. Les cinq niveaux de notation deviennent dans ce cas : 5- "Très confortable", 4- "Confortable", 3- "Légèrement inconfortable", 2- "Inconfortable" et 1- "Extrêmement inconfortable".

3.2.3 Évaluation de la qualité audio-visuelle

La plupart des méthodes d'évaluation subjective sont destinées à l'évaluation de contenus unimodaux. En effet, nous avons vu que la qualité de contenus audio peut être évaluée à l'aide de certaines normes (UIT-R BS 1116, 1534 et 1284), alors que d'autres normes sont applicables lorsqu'il s'agit de contenus vidéo 2D ou 3D (UIT-R BT 500, 1788 et 2021). Certaines normes suggèrent également des méthodes pour évaluer une modalité donnée (audio ou vidéo) dans un contexte audio-visuel : les norme UIT-R BS 1286 [ITU 1286, 1997] pour évaluer les systèmes audio en présence d'une image d'accompagnement et la norme UIT-T P910 [ITU P910, 2008] pour évaluer la qualité vidéo dans le cadre de systèmes multimédias tels que la visioconférence. La norme UIT-T P911 est la seule norme dédiée à l'évaluation subjective de la qualité audio-visuelle pour les applications multimédias [ITU P911, 1998].

Cette norme propose d'évaluer le critère unique de la qualité audio-visuelle globale qui peut être estimée grâce à quatre protocoles d'évaluation :

— la méthode à un seul stimulus SS (échelle de qualité),
— la méthode à double stimulus DSIS (échelle de dégradation),
— la méthode de comparaison SC (échelle de comparaison),
— la méthode d'évaluation continue à un seul stimulus SSCQE (échelle de qualité).

Les échelles de notation sont identiques à celles préconisées dans les normes présentées précédemment (voir figure 3.1). Il est cependant recommandé d'utiliser une échelle de qualité en neuf points si le niveau de discrimination entre les conditions d'essai est faible.

Les conditions d'écoute et de visualisation sont également spécifiées dans la norme UIT-T P911 (luminosité de la salle et de l'écran, résolution et contraste de l'écran, distance du participant au dispositif de diffusion, niveau de bruit de fond, niveau d'écoute, temps de réverbération de la salle, etc.).

3.2.4 Limitations des méthodes d'évaluation actuelles

Les méthodes d'évaluation subjective de la qualité de contenus audio, vidéo, ou audio-visuels comportent certaines limitations.

Les échelles de notation

Les méthodes d'évaluation subjective décrites précédemment recommandent l'utilisation d'échelles de notation telles que l'échelle de dégradation, de qualité et de comparaison. Ces échelles sont généralement continues, ce qui laisse la liberté aux participants de placer le curseur de notation à n'importe quel endroit entre les deux extrémités. Différentes catégories représentant les niveaux de qualité (également appelés items) sont réparties de manière régulière le long de l'échelle. Dans le cas de l'échelle de qualité par exemple, les cinq niveaux de qualité identifiés sont "Excellent", "Bon", "Assez bon", "Médiocre" et "Mauvais". Certaines études ont montré que l'uniformisation de ces items au niveau international peut générer un biais [Jones and McManus, 1986, Mullin et al., 2001, Zielinski et al., 2007]. En effet, les différences culturelles et linguistiques peuvent entraîner une compréhension des items changeante d'un pays à l'autre. Zielinski, par exemple, a comparé les résultats issus d'expériences menées dans plusieurs pays et a montré que la traduction des niveaux de qualité utilisés peut s'écarter de leur définition initiale. La différence perçue entre deux items successifs ("Excellent" et "Bon", par exemple) varie suivant la langue utilisée et le pays, comme l'illustre la figure 3.3.

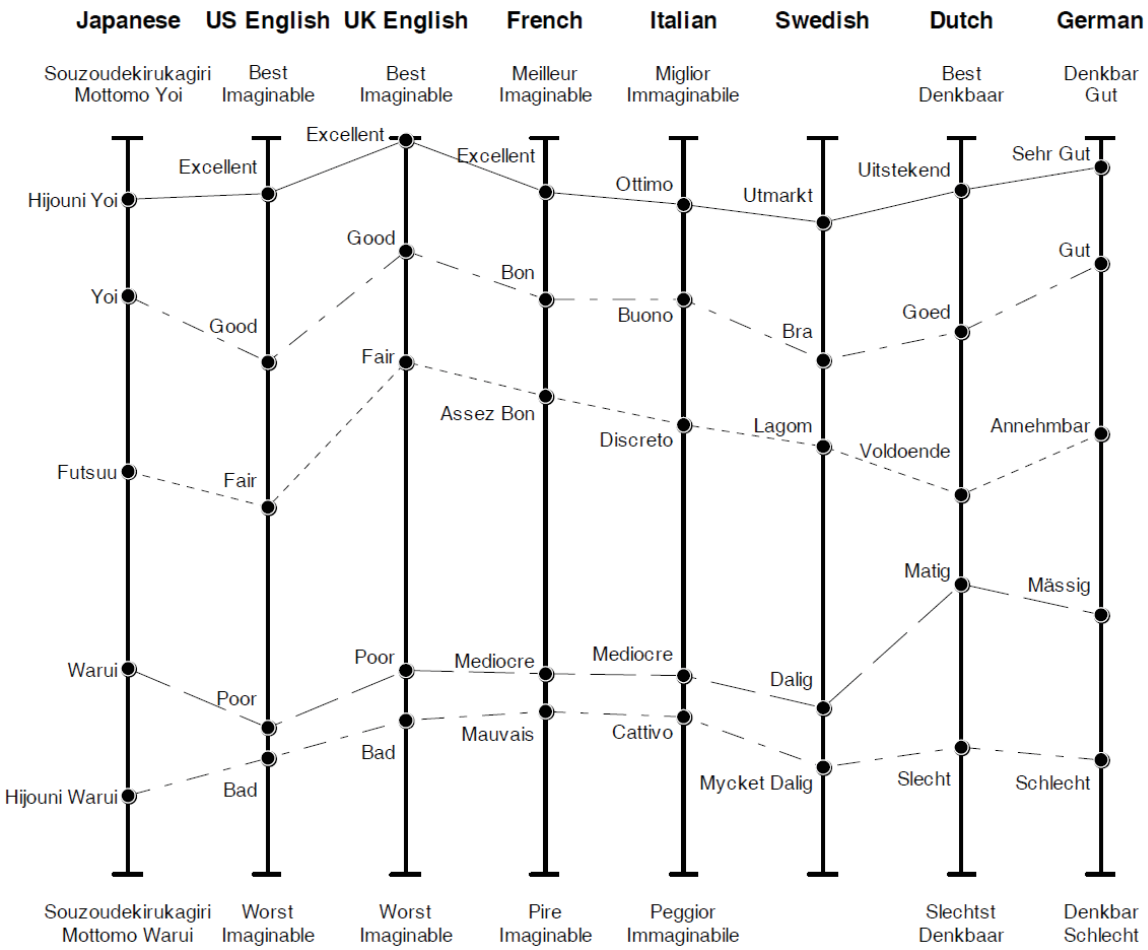


FIGURE 3.3 – *Différences linguistiques observées entre les items utilisés pour l'évaluation de la qualité (d'après [Zielinski et al., 2007]). Les participants devaient placer les adjectifs sur l'échelle, suivant leur signification, entre "meilleur imaginable" et "pire imaginable".*

Zielinski a également analysé la distribution des notes attribuées le long des échelles continues et a révélé des comportements différents suivant les sujets. En effet, certains participants utilisent toute la dynamique disponible de manière uniforme, alors que d'autres placent le curseur uniquement en face des étiquettes de niveau. De manière générale, que ce soit pour l'échelle de qualité ou pour l'échelle de dégradation, les participants placent plus fréquemment le curseur en face des étiquettes qu'aux autres points de l'échelle. Les points situés à mi-chemin entre deux étiquettes sont également préférés.

Méthodes d'évaluation pour les sons spatialisés

Les méthodes actuelles d'évaluation de la qualité audio utilisent majoritairement la "qualité audio de base" comme critère d'évaluation. La "qualité audio de base" est définie comme "la caractéristique unique et globale qui sert à estimer toute différence décelée entre la référence et l'objet à évaluer" [ITU 1534, 2014]. Cependant, certaines méthodes proposent des caractéristiques supplémentaires, applicables à des systèmes stéréophoniques et multicanaux. Pour l'évaluation des systèmes stéréophoniques, l'attribut supplémentaire proposé par les recommandations est la "qualité d'image stéréophonique". Celle-ci est associée à la différence entre la référence et le son à évaluer en termes d'emplacement des images sonores, d'impression de profondeur et de présence de l'évènement audio. Pour l'évaluation des systèmes multicanaux, les recommandations proposent deux attributs supplémentaires : "la qualité frontale de l'image", associée à la localisation des sources sonores frontales, et la "qualité d'impression ambiophonique", associée à une impression d'espace, à l'ambiance ou à des effets spéciaux directionnels particuliers. Il faut préciser que ces paramètres supplémentaires sont très rarement utilisés. Une raison tangible est la mise en garde présente dans les recommandations sur le manque de recul par rapport à l'utilisation de ces critères. Il y est notamment mentionné que "ces critères n'ont pas fait l'objet de recherches suffisantes" [ITU 1534, 2014]. D'autre part, les définitions de ces paramètres sont peut-être trop approximatives pour pouvoir les utiliser convenablement.

Dans le contexte de l'évaluation de sons spatialisés, il serait cependant intéressant de prendre en compte d'autres critères plus spécifiques que la "qualité audio de base", relatifs à la dimension spatiale de la scène sonore par exemple. De nombreuses études ont été menées dans le but de caractériser la qualité perçue des sons grâce à l'identification d'attributs spécifiques [Letowski, 1989, Bech, 1999]. Afin d'extraire ces critères, différentes méthodes d'élicitation ont été utilisées telles que l'analyse multidimensionnelle (MDS pour *Multi-Dimensional Scaling*) [Bech and Zacharov, 2006, Rumsey, 1998, Susini et al., 1999], la méthode "grille-répertoire" (RGT pour *Repertory Grid Technique*) [Berg and Rumsey, 1999, Berg, 2005], ou encore la méthode *Perceptual Structure Analysis* (PSA) [Choisel and Wickelmaier, 2006]. Toutes ces études fournissent des critères caractéristiques de la qualité sonore tels que le bruit de fond, la brillance, l'immersion, la coloration, la profondeur, l'enveloppement, ou encore la précision. La prise en compte de ces attributs dans un test d'écoute étant prohibitif, des études ont été menées pour regrouper ces attributs par catégorie [Letowski, 1989, Berg and Rumsey, 2003, Le Bagousse et al., 2010]. Les résultats de ces études tendent vers une distinction des aspects liés au timbre, à l'espace et parfois aux défauts techniques présents dans le signal sonore. D'autres travaux vont même jusqu'à affirmer que la qualité audio de base serait déterminée à 70 % par les

attributs liés au timbre (brillance, coloration, etc.) et à 30 % par les attributs liés à la spatialisation (immersion, profondeur, etc.) [Rumsey et al., 2005, Marins et al., 2008]. Dans le cas de l'évaluation de sons spatialisés, une méthode d'évaluation multi-critères prenant en compte ces différentes catégories d'attributs pourrait être une alternative aux méthodologies normalisées.

Méthodes d'évaluation pour les contenus audio-visuels 3D

Bien qu'il existe depuis peu une norme relative à l'évaluation de la qualité de vidéos 3D [ITU 2021, 2012], aucune recommandation n'est actuellement disponible concernant les contenus audio-visuels impliquant des vidéos 3D. Compte tenu des limitations des normes existantes et des études menées dans ce domaine, l'utilisation de plusieurs critères de notation semble pertinente pour juger la qualité de contenus multimodaux. Il est en effet important de pouvoir identifier l'impact des pertes de qualité audio ou vidéo sur l'expérience audio-visuelle 3D globale. Or, une évaluation suivant un critère unique (la "qualité audio-visuelle globale", par exemple) ne permettrait pas d'obtenir ce type d'informations. Cet aspect est pris en compte dans l'expérience subjective présentée dans le suite de ce chapitre. En effet, une méthodologie de test basée sur une approche multi-critères est utilisée pour évaluer l'expérience audio-visuelle associée à la présentation de contenus 3D.

Limitation de la qualité d'expérience à la qualité perçue des contenus

L'Union Internationale des Télécommunications définit la qualité d'expérience (QoE) comme l'acceptabilité d'un service tel que subjectivement perçu par l'utilisateur final [ITU P10, 2008]. Cependant, pour la plupart des applications multimédias, l'évaluation de la QoE est généralement limitée à la qualité des signaux audio et/ou vidéo perçue par l'utilisateur. Or, l'évaluation de la qualité des médias ne permet pas à elle seule de déterminer l'acceptabilité d'un service. En effet, le contexte d'utilisation et les attentes des utilisateurs (en termes de plaisir, de confort, etc.) sont autant d'aspects qui peuvent potentiellement affecter la qualité d'expérience. Notons que seules les méthodologies d'évaluation dédiées aux vidéos 3D semblent s'intéresser à ces notions en prenant en compte le critère de gêne/confort de visualisation. Dans la même idée, l'évaluation rigoureuse de la QoE associée à des services audio spatialisés ou audio-visuels 3D nécessiterait d'adapter les méthodologies existantes en prenant en compte des critères supplémentaires tels que le confort ou l'immersion, par exemple.

3.3 Évaluation multi-critères de l'expérience audio-visuelle 3D actuelle

Cette première expérimentation a pour but d'évaluer la qualité d'expérience audio-visuelle 3D proposée par les technologies de reproduction, audio et vidéo 3D, utilisées aujourd'hui. Cette étude de la situation existante est primordiale puisqu'elle va permettre de comprendre les limitations des systèmes de diffusion sonore actuellement associés à la présentation de vidéos 3D.

3.3.1 Contexte de l'étude

Il a été montré dans le chapitre 2 que le format de restitution sonore multicanal 5.1 est généralement utilisé dans les installations audio-visuelles destinées au grand public (cinéma, télévision, etc.). Ce monopole du format 5.1 concerne aussi bien les contenus vidéo 2D que 3D. La diffusion de contenus sonores mixés au format 5.1 peut être réalisée à l'aide de différents types d'équipements disponibles dans le commerce. La solution de diffusion la plus naturelle consiste à reproduire les 5 canaux d'origine à l'aide de 5 haut-parleurs, disposés conformément à la recommandation UIT-R BS 775 [ITU 775, 2012] (voir figure 2.2). Une autre solution consiste à adapter les 5 canaux audio d'origine au système de diffusion constitué de N haut-parleurs. Cette phase d'adaptation est totalement transparente pour l'utilisateur puisqu'elle est réalisée au sein des équipements grâce aux méthodes d'*up-mixing* (lorsque $N > 5$) ou de *down-mixing* (lorsque $N < 5$). Cette solution permet par exemple de diffuser une source initialement mixée en 5.1 sur un système de haut-parleurs stéréophonique constitué de deux haut-parleurs. Ce *down-mix* d'un signal 5.1 vers le 2.0 est également compatible avec une écoute au casque. Enfin, les barres sonores sont disponibles sur le marché depuis la fin des années 2000 et représentent des équipements compacts parfois capables de restituer le signal sonore 5.1 grâce aux techniques de *beam-forming*. Cette technique permet de créer des faisceaux sonores suivant différentes directions. Certains modèles de barre sonore utilisent cette technique pour récréer des haut-parleurs virtuels placés à différentes positions, en exploitant notamment les réflexions des ondes acoustiques sur les parois de la salle de diffusion.

Différents moyens de diffusion sonore capables de restituer un signal audio initialement mixé en 5.1 sont donc actuellement accessibles au grand public (système 5.1, casque, barre sonore). Ces équipements impactent inévitablement l'expérience sonore des spectateurs, mais qu'en est-il pour l'expérience audio-visuelle 3D ? Le but de cette première expérimentation est de mesurer l'impact potentiel du système de restitution sonore sur la perception audio, visuelle et audio-visuelle 3D. L'idée est également de déterminer si, parmi les trois systèmes étudiés, l'un d'eux est plus adapté pour accompagner des vidéos 3D.

Pour ce faire, 15 séquences audio-visuelles 3D mixées en 5.1 sont présentées, grâce à trois systèmes de reproduction sonore (système 5.1, casque, barre sonore), puis évaluées suivant une approche multi-critères par un panel de testeurs.

3.3.2 Environnement expérimental

Les séquences vidéo 3D sont présentées sur un écran LCD stéréoscopique de 47 pouces LG 47 LW5500 offrant une résolution de 1920 par 1080 pixels. Les participants doivent porter des lunettes 3D polarisées durant toute la durée du test afin de bénéficier des effets de relief visuel. La technologie 3D passive a été préférée à la solution active car les lunettes 3D sont plus légères et donc plus confortables. De plus, nous souhaitons éviter un phénomène de battements qui pourrait être causé par l'interaction entre la fréquence d'obturation des lunettes actives et la présence de sources lumineuses dans la salle d'expérimentations (voir section 2.3.1).

Concernant la reproduction sonore, trois systèmes de restitution sont utilisés pour reproduire le flux audio d'origine mixé au format 5.1 : un système multicanal 5.1 Genelec 8040A bi-amplifié, une barre de son Yamaha YSP-1 et un casque audio ouvert AudioTechnica ATH-AD700. Les trois systèmes de diffusion sont pilotés par une carte son externe Terratec Phase 26 USB.

L'expérience est réalisée dans une salle d'écoute dont les parois bénéficient d'un traitement acoustique. Dans cette pièce, le temps de réverbération Tr_{60} est d'environ 350 ms, le niveau de bruit de fond est inférieur à 30 $dB(A)$, et la luminosité de la pièce est inférieure à 20 lux, conformément à la recommandation UIT-T P911 [ITU P911, 1998]. La distance de visualisation recommandée dans le cas d'un écran HD (1920 par 1080 pixels) est d'environ trois fois la hauteur de l'écran [ITU 500, 2012, ITU 2021, 2012]. Compte tenu des dimensions de l'écran utilisé, le spectateur est assis à une distance de 1,85 mètres du téléviseur 3D. La barre sonore est placée en dessous de l'écran (voir figure 3.4), et le système multicanal 5.1 est installé autour du spectateur, conformément à la norme UIT-R BS 775 [ITU 775, 2012], à une distance de 1,90 mètres du *sweet spot* (voir section 2.2.1).

FIGURE 3.4 – *Photographie représentant l'emplacement de la télévision LG (47 pouces de diagonale), la barre de son Yamaha placée sous l'écran ainsi que les trois haut-parleurs frontaux du système multicanal 5.1 Genelec utilisé.*

3.3.3 Stimuli

Le corpus de stimuli est constitué de 15 séquences audio-visuelles extraites d'un documentaire 3D sur le boxeur français Jean-Marc Mormeck [Drhey, 2009]. Chaque extrait dure entre 11 et 17 secondes. Les séquences ont été choisies pour couvrir un large éventail de possibilités quant aux relations entre le contenu sonore (parole, musique, bruits d'environnement, ambiances, etc.), et le contenu visuel (intérieur/extérieur, nombre de personnages, dynamique, etc.). Les caractéristiques des séquences sont présentées dans le tableau 3.1.

Séq.	Lieu	Contenu	Musique	Voix	Effets sonores	Boîte scénique (px)	Description
1	Extérieur	Dynamique	Oui	-	Oiseaux*	43	Footing (caméra épaule)
2	Extérieur	Dynamique	Oui	Respiration	Oiseaux * Bruits de pas	16	Footing (travelling)
3	Extérieur	Dynamique	Oui	Respiration	Oiseaux * Bruits de pas	41	Titre et footing
4	Intérieur	Dynamique	Oui	-	Bruit de foule* Coup de poing	34	Installation des sacs de frappe
5	Intérieur	Statique	Oui	Oui	-	33	Interview
6	Intérieur	Statique	Oui	-	Coup de poing	34	Entraînement sur poire de vitesse
7	Intérieur	Statique	Oui	Voiceover	-	37	Musculation
8	Intérieur	Statique	Oui	Voiceover	-	40	Echauffement
9	Intérieur	Statique	Oui	Voiceover	-	30	Discussion avec entraîneur
10	Intérieur	Dynamique	Oui	-	-	37	Entraînement au combat
11	Intérieur	Dynamique	-	-	Bruit de foule* Coup de poing	48	Combat
12	Intérieur	Dynamique	Oui	-	-	39	Phase de récupération
13	Intérieur	Statique	Oui	Voiceover	-	21	Phase de récupération (ralenti)
14	Intérieur	Dynamique	Oui	-	Bruit de foule*	54	Installation des sacs de frappe, crédits
15	Intérieur	Statique	-	Voiceover	-	51	Générique final

TABLE 3.1 – *Description des séquences de test sélectionnées.*

La boîte scénique décrit de manière objective la quantité de relief présent dans chaque extrait. Sa valeur est déterminée par la différence de disparité entre l'objet le plus proche et le plus lointain présent à l'image. Ces disparités sont mesurées en pixels grâce à un système StereoLabs PURE. Des effets sonores *surround* ont été ajoutés en post-production sur les séquences annotées (*) dans le tableau 3.1.

Les fichiers vidéo sont compressés à un débit moyen de 30 $Mbps$ par un codec H.264 et présentés à une fréquence de 25 images par seconde. Aucune compression n'est affectée aux fichiers audio (PCM) qui sont diffusés à un débit de 6912 $kbps$ dans le cas des fichiers 5.1 et 1536 $kbps$ pour les fichiers stéréo ayant subi la phase de *down-mixing*[1]. La fréquence d'échantillonnage des signaux audio est de 48 kHz.

1. *Down-mix* réalisé d'après les facteurs définis dans la norme UIT-R BS 775.

3.3.4 Constitution du panel de participants

Le panel de participants est constitué de 30 personnes (21 femmes et 9 hommes) dont la moyenne d'âge est d'environ 33 ans. Vingt-huit de ces participants ont l'habitude de réaliser des tests d'écoute, mais aucun d'entre eux n'a déjà réalisé de test subjectif portant sur la qualité de contenus audio-visuels 3D. Les testeurs sont rémunérés pour leur participation et sont considérés comme présentant un profil de type "grand public" dans le contexte de cette étude. En effet, aucun d'entre eux n'exerce une activité professionnelle ayant un lien avec le domaine de l'audio ou de la vidéo.

3.3.5 Protocole expérimental

L'expérience est divisée en trois sessions. Chaque session est dédiée à un système de reproduction sonore : le système multicanal 5.1 (1), le casque (2) ou la barre sonore (3). L'ordre des sessions varie suivant les participants de manière à ce que chaque combinaison possible (présentation des systèmes 1/2/3, ou 1/3/2, ou 2/1/3, etc.) soit présentée au même nombre de personnes. Il est ainsi possible de s'affranchir d'un effet potentiel lié à l'ordre de présentation des systèmes de reproduction étudiés. Au sein d'une session, les séquences sont également présentées dans un ordre aléatoire à chaque participant.

Lors de chaque session de test, les 15 séquences audio-visuelles sont présentées trois fois, de sorte que les participants puissent se concentrer successivement sur les propriétés sonores, visuelles, puis audio-visuelles des extraits :

- Durant la première présentation, les spectateurs ont pour consigne d'évaluer le **degré de profondeur visuelle** et le **confort de visualisation** (propriétés vidéo).
- Durant la deuxième présentation, les spectateurs jugent la **degré de spatialisation sonore** et le **confort d'écoute** (propriétés audio).
- Enfin, lors de la troisième présentation, les spectateurs doivent considérer les séquences audio-visuelles dans leur globalité et évaluer le **degré de cohérence** entre le son et l'image, ainsi que le **degré d'immersion** dans la scène (propriétés audio-visuelles).

Après chaque visualisation de séquence, les participants jugent les deux critères associés aux propriétés audio, vidéo ou audio-visuelles. La notation s'effectue sur des échelles discrètes à 5 points pour tous les critères. Les extrémités de ces échelles portent les mentions "inconfortable / confortable" pour les critères relatifs au confort de visualisation et au confort d'écoute, et "faible / fort" pour les autres critères (degré de profondeur visuelle, degré de spatialisation sonore, cohérence et immersion audio-visuelle).

Avant la phase de test, une phase de familiarisation est réalisée sur 4 séquences extraites du même documentaire 3D. Cette phase d'entraînement permet aux participants de se familiariser avec l'interface de test et le protocole d'évaluation.

3.3.6 Résultats

La durée totale de test est en moyenne de 92 minutes (pauses comprises). Pour chaque critère évalué, une analyse de la variance (ou ANOVA pour *ANalysis Of VAriance*) est effectuée sur l'ensemble des jugements en considérant deux facteurs : le système de restitution sonore et la séquence évaluée. Les coefficients de corrélation intercritères sont également calculés.

Critères liés à la vidéo 3D

Concernant le **degré de profondeur visuelle**, la figure 3.5 illustre les notes moyennes et les intervalles de confiance à 95 % associés obtenus pour chaque système de restitution sonore, et chaque séquence. Un effet de séquence est révélé par l'analyse ANOVA ($F(14,406)=11.27$, $p<0.001$), ce qui indique que les spectateurs ont perçu des différences de profondeur visuelle entre les différentes séquences présentées. Le système de restitution sonore ne semble pas avoir d'effet sur le degré de profondeur visuelle perçu ($F(2,58)=0.48$, $p=0.62$). Cette absence d'effet semble commune à toutes les séquences puisqu'aucune interaction n'est décelée entre le système de restitution évalué et la séquence ($F(28,812)=0.69$, $p=0.88$).

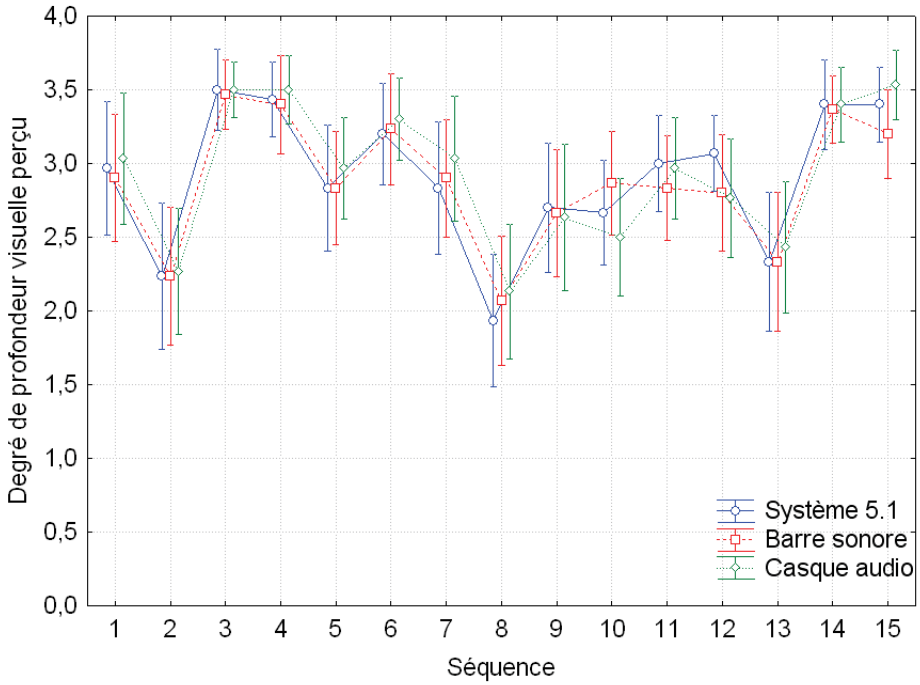


FIGURE 3.5 – *Notes moyennes et intervalles de confiance à 95 % associés pour le critère "Profondeur visuelle". Les jugements sont notés entre 0, pour un degré de profondeur visuelle "faible", et 4, pour un degré de profondeur visuelle "fort".*

Nous proposons maintenant de comparer les notes de profondeur visuelle fournies par les participants à un indice de profondeur objectif : la boîte scénique. La figure 3.6 représente le degré de profondeur visuelle perçu en fonction de la taille de la boîte scénique de chaque séquence (voir tableau 3.1).

Une relation linéaire peut être établie entre les notes de profondeur visuelle perçue et l'étendue de la boîte scénique pour 11 séquences avec un coefficient de détermination R^2 de 0.89 (losanges bleu sur la figure 3.6). Cependant, l'étendue de la boîte scénique ne semble pas constituer un critère suffisant pour expliquer tous les jugements de profondeur visuelle. En effet, la profondeur visuelle perçue est plus importante que prévue par ce modèle linéaire pour les séquences 3, 4 et 6 (points rouge sur la figure 3.6). Une hypothèse est que, pour ces séquences particulières, la perception du relief est augmentée par la présence d'indices visuels

supplémentaires. Il est important de rappeler que l'étendue de la boîte scénique est déterminée pour chaque séquence, à partir des disparités de l'objet le plus proche et le plus lointain. Or, il a été montré dans la section 1.3.1 que de nombreux indices visuels contribuent à la perception du relief. Dans les séquences 3, 4 et 6, des indices monoculaires tels que la perspective linéaire et la taille relative des objets sont présents et peuvent potentiellement augmenter la sensation de relief visuel. De plus, pour la séquence numéro 3, l'apparition dynamique du titre en jaillissement (devant l'écran) peut accentuer cet effet.

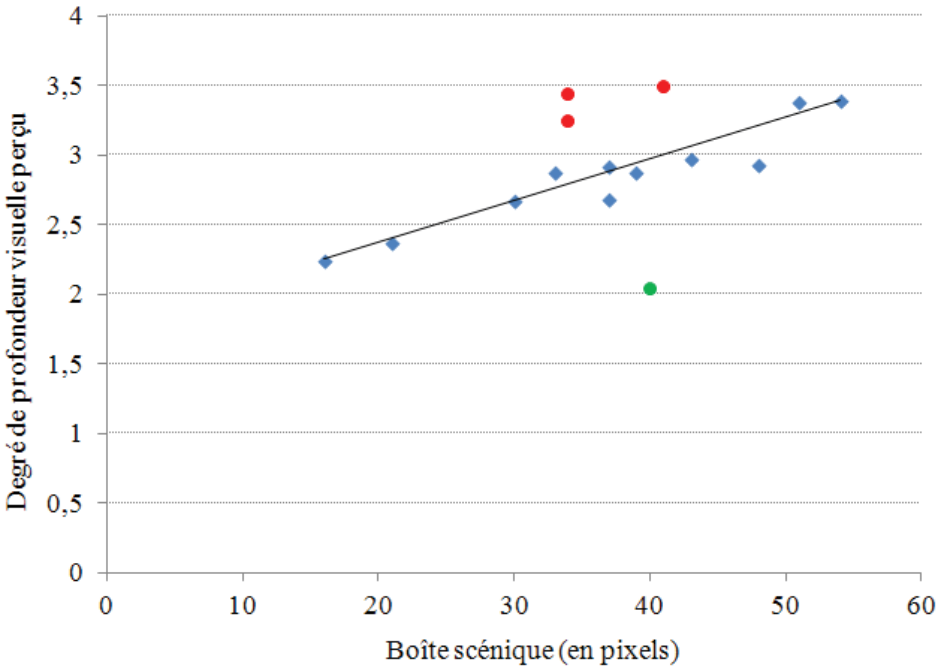


FIGURE 3.6 – *Notes moyennes de profondeur visuelle perçue en fonction de l'étendue de la boîte scénique mesurée.*

Dans le but d'illustrer ces hypothèses, des captures d'écran des séquences 3 et 4 sont présentées sur la figure 3.7. La prédiction de la profondeur visuelle perçue est également limitée pour la séquence 8 (point vert sur la figure 3.6). Pour cette séquence, la perception du relief visuel est inférieure à la valeur estimée par le modèle de prédiction basé sur l'étendue de la boîte scénique. Plusieurs participants ont déclaré que le relief de cette séquence était perçu comme "artificiel", ce qui peut expliquer une dégradation de la perception du relief. Cet effet de relief artificiel est causé par une perception "aplatie" des objets en 3D ou des différents plans d'une image. Ce phénomène, également appelé *cardboard effect*, serait causé par une répartition des objets visuels dans la scène et par des paramètres de captation vidéo 3D mal adaptés [Yamanoue et al., 2000, Yamanoue et al., 2006].

FIGURE 3.7 – *Capture d'écran des séquences 3 (à gauche) et 4 (à droite). Dans la séquence 3, la profondeur visuelle est augmentée par la perspective linéaire du chemin sur lequel court le personnage, ainsi que l'apparition dynamique du titre. Dans la séquence 4, les indices visuels de profondeur ajoutés sont la taille relative des sacs de frappe et l'alignement des sacs (perspective linéaire).*

Le second critère d'évaluation propre à l'expérience visuelle est le **confort de visualisation**. La figure 3.8 présente les notes moyennes ainsi que les intervalles à 95 % pour ce critère.

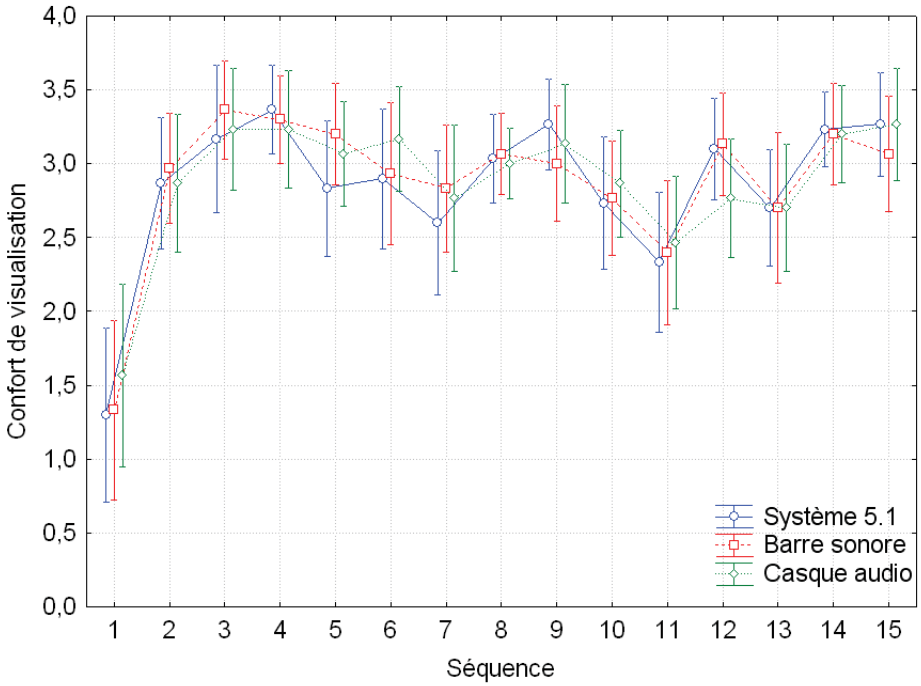


FIGURE 3.8 – *Notes moyennes et intervalles de confiance à 95 % associés pour le critère "Confort de visualisation". Les jugements sont notés entre 0, pour une séquence jugée "inconfortable", et 4, pour une séquence "confortable".*

Comme dans le cas du degré de profondeur visuelle, le confort de visualisation dépend exclusivement de la séquence évaluée ($F(14,406)=10.21$, $p<0.001$). Cet effet est principalement porté par la première séquence qui est perçue comme étant inconfortable pour tous les systèmes

de restitution sonore étudiés. Cette séquence présente la particularité d'être filmée en "caméra épaule", ce qui induit des mouvements de caméra marqués en comparaison des autres séquences. De plus, l'apparition du titre en jaillissement occasionne des disparités négatives importantes. La combinaison de ces deux phénomènes explique certainement le faible confort de visualisation associé à cette séquence. Le système de restitution sonore ne semble pas avoir d'impact sur le confort de visualisation ($F(2,58)=0.24$, $p=0.78$).

Critères liés à l'audio

Lors de la deuxième présentation des séquences audio-visuelles, les spectateurs se concentrent sur les propriétés audio des conditions présentées et évaluent le **degré de spatialisation sonore** (figure 3.9) et le **confort d'écoute** (figure 3.10).

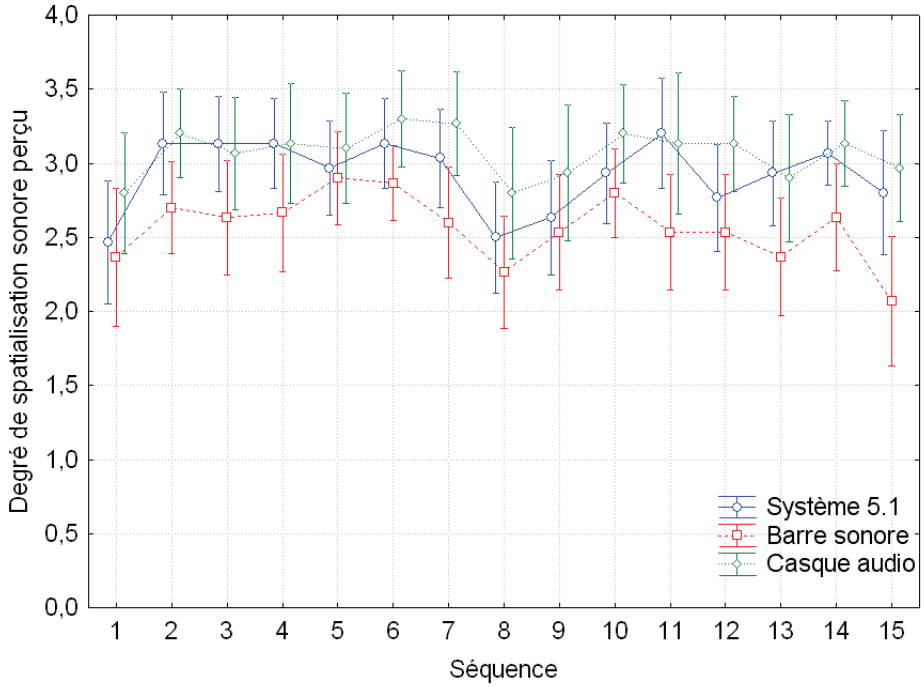


FIGURE 3.9 – *Notes moyennes et intervalles de confiance à 95 % associés pour le critère "Spatialisation sonore". Les jugements sont notés entre 0, pour un degré de spatialisation "faible", et 4, pour un degré de spatialisation "fort".*

L'analyse ANOVA révèle que le degré de spatialisation sonore est sensible aux séquences évaluées ($F(14,406)=3.29$, $p<0.001$), mais varie également de manière significative en fonction du système de restitution sonore utilisé ($F(2,58)= 9.88$, $p<0.001$). La figure 3.9 illustre cet effet du système de reproduction sonore et fait apparaître une hiérarchisation entre les trois systèmes étudiés. En effet, la diffusion au casque semble procurer une sensation de spatialisation sonore légèrement plus marquée que pour le système multicanal 5.1, alors que la barre sonore est le système de diffusion qui offre le moins d'effets de spatialisation. D'un point de vue statistique, cette observation est valable pour toutes les séquences puisqu'aucune interaction n'est révélée entre le système de restitution et les séquences ($F(28,812)=0.88$, $p=0.64$).

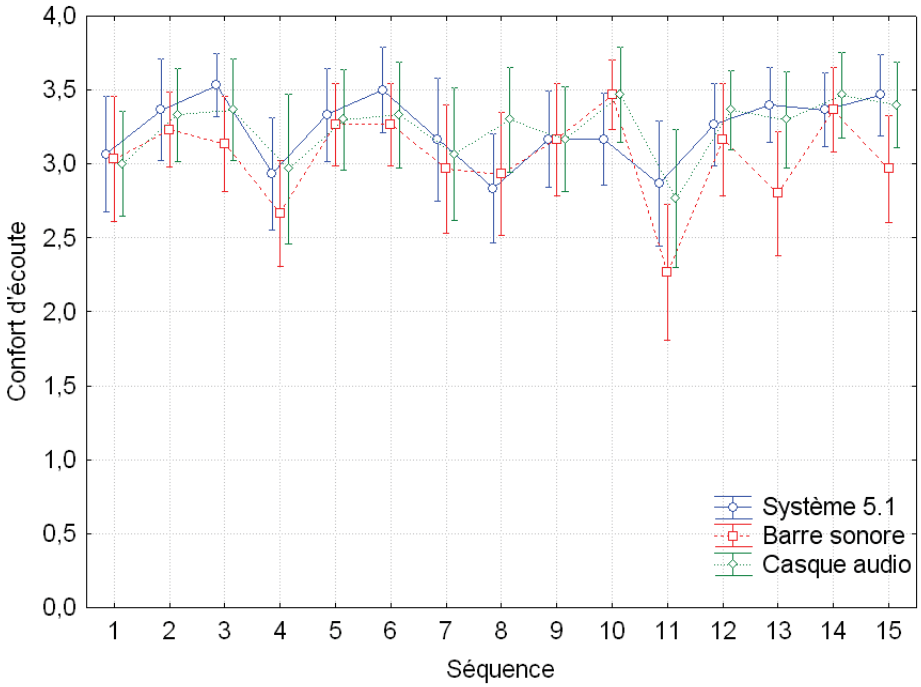


FIGURE 3.10 – *Notes moyennes et intervalles de confiance à 95 % associés pour le critère "Confort d'écoute". Les jugements sont notés entre 0, pour une écoute "inconfortable", et 4, pour une écoute "confortable".*

Le confort d'écoute semble, quant à lui, comparable, quel que soit le système de restitution utilisé ($F(2,58)=2.54$, $p=0.09$). En effet, la figure 3.10 montre que les séquences sont généralement perçues comme étant confortables à écouter (notes comprises entre 3 et 3.5), indépendamment du système de diffusion sonore employé. Ce résultat suggère que les participants n'ont pas été gênés par le port simultané du casque audio et des lunettes 3D. Il est à noter que le choix du casque (AudioTechnica ATH-AD700) joue probablement un rôle important dans ce jugement. En effet, la pression exercée par ce casque sur la tête des auditeurs est faible par rapport à certains modèles. Les résultats obtenus pour ce critère seraient probablement différents si un autre casque audio avait été utilisé.

Enfin, l'analyse ANOVA montre que le confort d'écoute varie suivant la séquence présentée ($F(14,406)=4.76$, $p<0.001$). Les séquences 4 et 11 sont, par exemple, perçues comme étant moins confortables. Une explication pourrait venir du contenu sonore propre à ces deux séquences. En effet, le tableau 3.1 révèle que des effets sonores communs sont présents dans les séquences 4 et 11 (bruit de foule et coup de poing). Il est probable que le spectre large bande, les bruits d'impact et les cris de foule très présents dans ces séquences soient perçus comme étant trop agressifs et engendrent de l'inconfort.

Critères audio-visuels

Lors de la dernière présentation des séquences, les participants évaluent les critères audio-visuels relatifs à la **cohérence entre le son et l'image** d'une part (figure 3.11), et la **sensation d'immersion** dans la scène d'autre part (figure 3.12).

L'analyse ANOVA révèle la présence d'un effet significatif (bien que très faible) du système de reproduction sonore sur le jugement de cohérence audio-visuelle ($F(2,58)=4.41$, $p<0.05$). Un test *post-hoc* de type Bonferroni révèle que cet effet est significatif uniquement entre le casque audio et la barre sonore. La figure 3.11 montre en effet que, pour la plupart des séquences, la cohérence entre le son et l'image est légèrement plus importante dans le cas de l'écoute au casque. Cependant, les jugements de cohérence semblent davantage impactés par le contenu des séquences que par le système de restitution ($F(14,406)=4.44$, $p<0.001$).

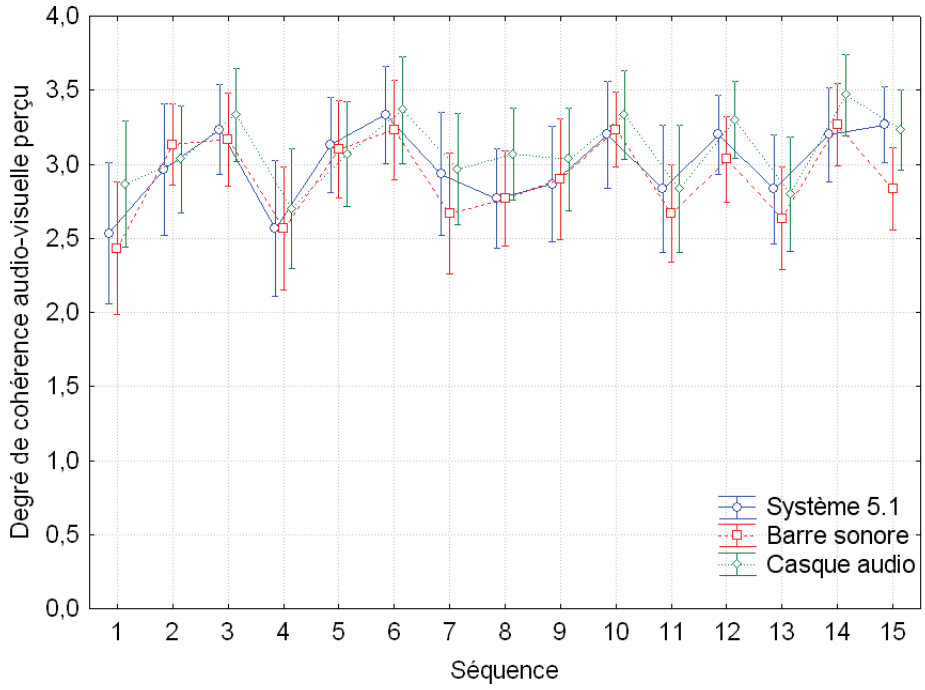


FIGURE 3.11 – *Notes moyennes et intervalles de confiance à 95 % associés pour le critère "Cohérence audio-visuelle". Les jugements sont notés entre 0, pour un degré de cohérence "faible", et 4, pour un degré de cohérence "fort".*

Le degré d'immersion dans la scène audio-visuelle est également fortement dépendant de la séquence ($F(14,406)=4.28$, $p<0.001$), sans pour autant présenter de différences significatives entre les trois systèmes de restitution sonore ($F(2,58)=2.69$, $p=0.08$).

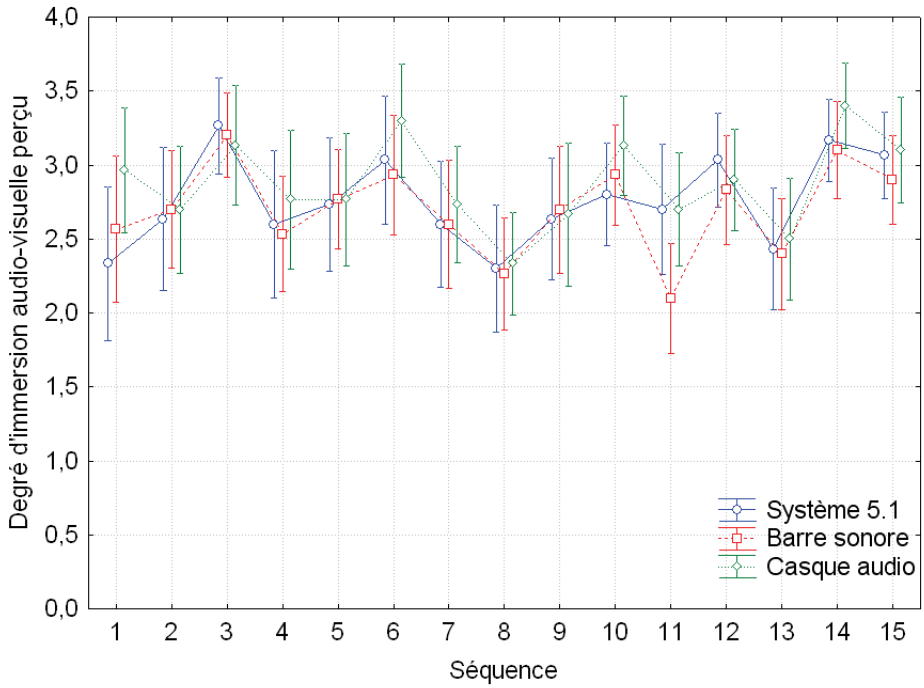


FIGURE 3.12 – *Notes moyennes et intervalles de confiance à 95 % associés pour le critère "Immersion audio-visuelle". Les jugements sont notés entre 0, pour un degré d'immersion "faible", et 4, pour un degré d'immersion "fort".*

Les résultats présentés précédemment (pour les critères vidéo et audio) ne semblent pas expliquer de manière triviale les jugements de cohérence et d'immersion audio-visuelle. Une analyse des corrélations existantes entre les critères d'évaluation a donc été réalisée. Le tableau 3.2 présente les coefficients de corrélation calculés. Les critères V1 et V2 représentent respectivement le degré de profondeur visuelle et le confort de visualisation. Les critères A1 et A2 sont les critères propres à la spatialisation sonore et au confort d'écoute. Enfin, AV1 représente la cohérence audio-visuelle, et AV2 est le degré d'immersion dans la scène audio-visuelle.

Critères d'évaluation		**Vidéo**		**Audio**		**Audio-Visuels**	
		V1	**V2**	**A1**	**A2**	**AV1**	**AV2**
Vidéo	**V1**	1	0,41	0,23	0,19	0,36	0,46
	V2	-	1	0,20	0,25	0,38	0,46
Audio	**A1**	-	-	1	0,62	0,46	0,43
	A2	-	-	-	1	0,57	0,49
Audio-Visuels	**AV1**	-	-	-	-	1	0,67
	AV2	-	-	-	-	-	1

TABLE 3.2 – *Coefficients de corrélation calculés sur les six critères d'évaluation.*

Cette analyse confirme l'absence de relation claire entre les critères. Certaines observations peuvent cependant être faites. La corrélation élevée entre les critères A1 et A2 (0,62) indique que les deux critères choisis pour rendre compte des propriétés audio des séquences ne sont pas orthogonaux. De même, un coefficient de corrélation entre les critères A2 et AV1 de 0,57 semble montrer que l'évaluation de la cohérence audio-visuelle est conditionnée par le jugement de confort d'écoute. Enfin, le critère d'immersion AV2 semble dépendre de tous les critères évalués (coefficients supérieurs à 0,43), tout en présentant une corrélation plus importante avec le critère AV1 relatif à la cohérence entre le son et l'image (0,67).

3.3.7 Discussions

Les résultats de cette expérience ont montré que l'expérience visuelle ne semble pas être impactée par le choix du système de restitution sonore (casque, système multicanal 5.1 ou barre sonore). Concernant l'impact potentiel du système audio sur l'expérience audio-visuelle, seule la notion de cohérence entre le son et l'image a été faiblement impactée par le système de diffusion ($F(2,58)=4.41$, $p<0.05$). Pour ce critère, la restitution au casque a été jugée comme légèrement plus cohérente en comparaison de la barre sonore (aucun effet significatif n'a été observé par rapport au système 5.1). Cet effet est à considérer avec prudence puisque l'analyse des corrélations réalisée entre les critères montre que la cohérence audio-visuelle est corrélée avec le confort d'écoute (0,57), et que ce dernier est potentiellement lié au modèle de casque utilisé lors de cette expérience. En d'autres termes, l'utilisation d'un casque d'écoute moins confortable pourrait diminuer le confort d'écoute associé à ce moyen de restitution, et peut-être dégrader la cohérence audio-visuelle. Compte tenu de ces observations, il est difficile d'établir la prédisposition d'un des systèmes de diffusion sonore testés pour accompagner la vidéo 3D.

Le tableau 3.2 révèle cependant une corrélation importante entre l'immersion audio-visuelle et la cohérence entre le son et l'image. Une piste d'investigation pour améliorer l'immersion, et donc l'expérience audio-visuelle, consisterait à utiliser un système de restitution sonore capable de procurer un degré de cohérence audio-visuelle plus important. La notion de cohérence touche différents aspects : la cohérence sémantique entre les stimuli, la cohérence temporelle mais également la cohérence spatiale audio-visuelle. La question est donc de savoir si un système de restitution sonore est capable d'améliorer un de ces aspects. Le lien sémantique entre les stimuli ou la cohérence temporelle ne semblent pas être critiques dans l'expérience présentée précédemment, dans le sens où les contenus utilisés sont naturels (lien sémantique figé) et synchronisés dans le temps. Cependant, l'amélioration de la cohérence audio-visuelle spatiale semble être une piste intéressante dans le contexte spécifique de la vidéo 3D. En effet, dans la situation actuelle, les objets visuels sont perçus à différentes distances (devant ou derrière l'écran) grâce à la présentation stéréoscopique, alors que la scène sonore, qui accompagne ces objets visuels, offre uniquement des effets de latéralisation plus ou moins étendus. Les technologies de restitution sonore classiquement utilisées dans les applications audio-visuelles ne prennent pas en compte cette nouvelle dimension apportée par la vidéo 3D. Or, la différence de localisation perçue entre les stimuli audio et visuels suivant la distance pourrait tout à fait altérer la notion de cohérence audio-visuelle, et ainsi dégrader l'immersion dans la scène, ou plus généralement, l'expérience audio-visuelle 3D.

La suite de ce document a pour but de déterminer si un système de restitution sonore donné est capable de créer des objets sonores perçus comme étant placés à différentes distances. Cette problématique sera traitée dans le chapitre 5. Il s'agira également d'évaluer dans quelle mesure la distance d'objets sonores est perçue comme étant cohérente avec la distance d'objets visuels présentés en 3D (chapitre 6). Enfin, dans le dernier chapitre de ce travail, il sera question de vérifier si l'ajout d'informations sonores suivant la distance peut améliorer notre expérience audio-visuelle 3D.

Chapitre 4

Système de restitution audio-visuelle associant WFS et rendu stéréoscopique 3D

4.1 Introduction

Il a été montré, dans l'expérimentation du chapitre 3, que la sensation d'immersion audio-visuelle semble corrélée avec la notion de cohérence entre le son et l'image. Dans l'idée d'améliorer l'immersion des spectateurs dans les contenus vidéo 3D, nous avons décidé d'orienter ce travail vers la notion de cohérence spatiale audio-visuelle. Ce point semble en effet perfectible puisque les systèmes de restitution sonore et visuelle 3D actuellement utilisés dans les installations destinées au grand public sont susceptibles d'entraîner la perception d'incohérences, en termes de localisation en distance, entre les stimuli audio et visuels. L'hypothèse sous-jacente est que la restitution d'objets sonores suivant la distance pourrait augmenter la sensation de cohérence avec les images 3D, et ainsi améliorer l'immersion des spectateurs. Il en résulterait une meilleure expérience audio-visuelle 3D. Dans le but de vérifier ces hypothèses, différentes expériences subjectives nécessitant la diffusion de contenus audio-visuels doivent être réalisées. Il est tout d'abord nécessaire de disposer d'un système de restitution audio-visuelle 3D adapté aux besoins de nos expériences.

Dans ce chapitre, nous allons nous attacher, dans un premier temps, à présenter le système de diffusion audio-visuelle 3D utilisé pour réaliser les expériences décrites dans les chapitres suivants. Le choix de la technologie de restitution sonore aboutira à la proposition d'un système de restitution associant la technologie Wave Field Synthesis au rendu visuel stéréoscopique. Enfin, les performances du système WFS seront évaluées de manière objective par des mesures acoustiques.

4.2 Choix de la technologie de restitution sonore

De nombreuses technologies de restitution sonore ont été présentées dans le chapitre 2, parmi lesquelles les technologies multicanales, le VBAP, le binaural, l'ambisonique, et la Wave Field Synthesis. Il a été montré que ces technologies permettent la restitution de champs sonores dans le plan horizontal et/ou vertical. Cependant, ces technologies n'ont pas le même niveau de maturité et ne sont pas toutes capables de créer des objets sonores virtuels placés à différentes distances, tout en préservant la qualité de la restitution. Au regard des éléments présentés dans le chapitre 2, la Wave Field Synthesis figure parmi les technologies les plus prometteuses pour restituer la distance de sources sonores.

La Wave Field Synthesis (WFS) est une technique de reproduction sonore basée sur la reproduction du front d'onde acoustique d'une source virtuelle. Cette technologie permet *a priori* de restituer des sources sonores placées à différentes distances sur une zone d'écoute étendue. En effet, sa capacité à restituer l'effet de parallaxe acoustique a été démontrée dans de nombreuses études [Wierstorf et al., 2012, Rébillat, 2011, Renard, 2000]. La figure 4.1 illustre cette propriété à travers la localisation d'une source virtuelle réalisée à partir de différentes positions d'écoute. La restitution de la parallaxe acoustique permet ainsi de créer une scène sonore virtuelle stable (en azimut et en distance) et cohérente pour tous les auditeurs placés dans la zone d'écoute.

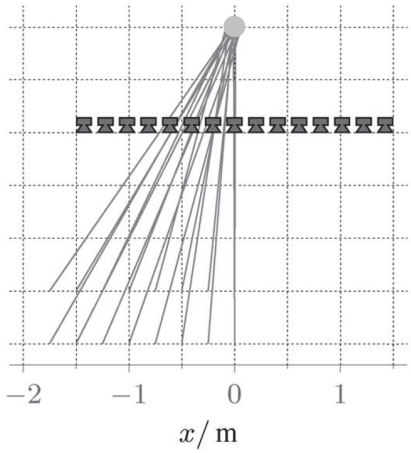


FIGURE 4.1 – *Direction perçue d'une source sonore virtuelle (cercle gris), évaluée à partir de 16 positions d'écoute, d'après [Wierstorf et al., 2012].*

Dans un contexte d'application audio-visuelle 3D, la Wave Field Synthesis présente un autre avantage. En effet, la WFS permet de créer des sources sonores virtuelles focalisées, c'est-à-dire au sein même de la zone d'écoute. Cette spécificité pourrait s'avérer particulièrement utile pour accompagner des images stéréoscopiques, car les éléments visuels 3D peuvent aussi être placés devant ou derrière l'écran. L'utilisation conjointe de la WFS et d'images stéréoscopiques pourrait ainsi offrir des objets audio-visuels spatialement cohérents, en amont et en aval de l'écran.

4.3 Description du système AV3D

Nous proposons dans cette section la description d'un système de restitution audio-visuelle associant la technologie Wave Field Synthesis à un rendu visuel stéréoscopique. Ce système sera utilisé dans les expérimentations présentées dans la suite de ce document ("système AV3D").

4.3.1 Restitution sonore

Les sources sonores virtuelles sont synthétisées par Wave Field Synthesis grâce à une rampe de 24 haut-parleurs Studer (voir figure 4.2). L'écart entre le centre de deux haut-parleurs consécutifs est de 9 cm. Cette distance entre les haut-parleurs entraîne l'apparition du phénomène de repliement spatial pour les fréquences supérieures à $f_{al} = 1890\ Hz$ (d'après l'équation (2.8) définie dans la section 2.2.5). Le dispositif WFS s'étend sur une largeur de $2,16\ m$ et est placé juste derrière l'écran de projection, à $1,1\ m$ de hauteur.

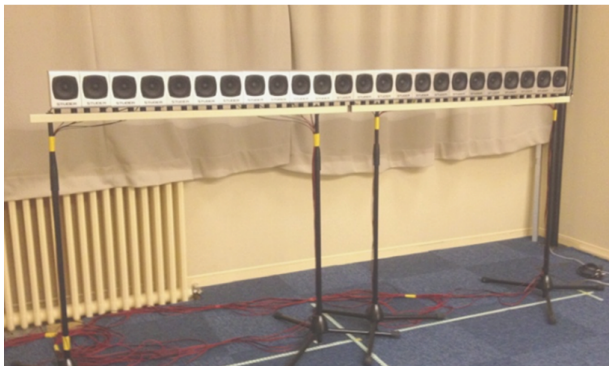


FIGURE 4.2 – *Système Wave Field Synthesis composé de 24 haut-parleurs espacés de 9 cm.*

Les réponses en fréquence des 24 haut-parleurs ont été mesurées au point d'écoute[1], puis égalisées afin de supprimer les différences de coloration spectrale propres à chaque haut-parleur. De plus, l'utilisation de la réponse en fréquence moyenne des 24 transducteurs mesurée au point d'écoute permet de s'affranchir de la directivité naturelle des haut-parleurs, et ainsi de simuler le rayonnement de sources omnidirectionnelles conformément au principe de la WFS. La synthèse du champ sonore par Wave Field Synthesis consiste à contrôler chacun des 24 haut-parleurs en gain et en retard. Les 24 signaux audio sont traités dans Max/MSP (filtres d'égalisation et implémentation WFS), puis transmis à une interface MOTU 24I/O via une carte son MOTU 424 PCIe. Un système d'amplification à 24 voies constitué de six amplificateurs de puissance Yamaha XM 4080 permet ensuite la restitution du champ sonore synthétisé.

4.3.2 Restitution visuelle 3D

La restitution visuelle 3D est assurée par un vidéoprojecteur 3D à technologie passive (LG CF3D), placé à 3 m d'un écran de projection. La projection passive nécessite l'utilisation d'une toile de projection métallisée afin de conserver la polarisation des images droite et gauche. Le

1. Mesures réalisées à l'aide d'un microphone omnidirectionnel.

port de lunettes 3D polarisées permet aux spectateurs de bénéficier des effets stéréoscopiques. La toile de projection est également micro-perforée pour limiter son effet de diffraction sur les ondes acoustiques émises par les haut-parleurs placés derrière la toile. L'image est projetée au format 16/9 et mesure 2 m de base sur $1,13$ m de hauteur. La distance de visualisation est fixée à 3 m, soit environ 3 fois la hauteur de l'écran comme recommandé par la norme UIT-R BT 500 [ITU 500, 2012].

Afin d'éviter une possible gêne visuelle lors de la restitution 3D, les disparités maximales à ne pas dépasser doivent être définies. Les objets virtuels doivent être placés dans une gamme de distance appelée "boîte scénique". Les limites de cette dernière sont déterminées avec le logiciel StereoCalculator sur la base du critère DoF (*Depth of Focus*) fixé à 0.2 [Chen et al., 2010, ITU 2021, 2012]. Compte tenu de la distance de visualisation (3 m) et de la taille de l'image, la restitution visuelle est confortable si les objets virtuels sont présentés entre $1,87$ m et $7,5$ m du spectateur. Ainsi, l'objet le plus proche équivaut à un jaillissement maximal de $1,13$ m devant l'écran, et le plus lointain peut se trouver jusqu'à $4,5$ m derrière l'écran.

Suivant les besoins expérimentaux, les contenus visuels 3D sont générés dans le moteur de jeu Unity 3D ou directement diffusés grâce au lecteur vidéo intégré de Max/MSP.

Les performances du vidéoprojecteur ont été évaluées de manière objective par différentes mesures optiques (colorimétrie, ghosting, gamma, temps de réponse, uniformité de luminosité sur l'écran, etc.), et semblent tout à fait satisfaisantes au regard des besoins expérimentaux.

La figure 4.3 illustre l'architecture générale du système de restitution audio-visuelle 3D proposé. Cette figure permet de se rendre compte des différents traitements réalisés afin de fournir à l'utilisateur des informations sonores et visuelles cohérentes.

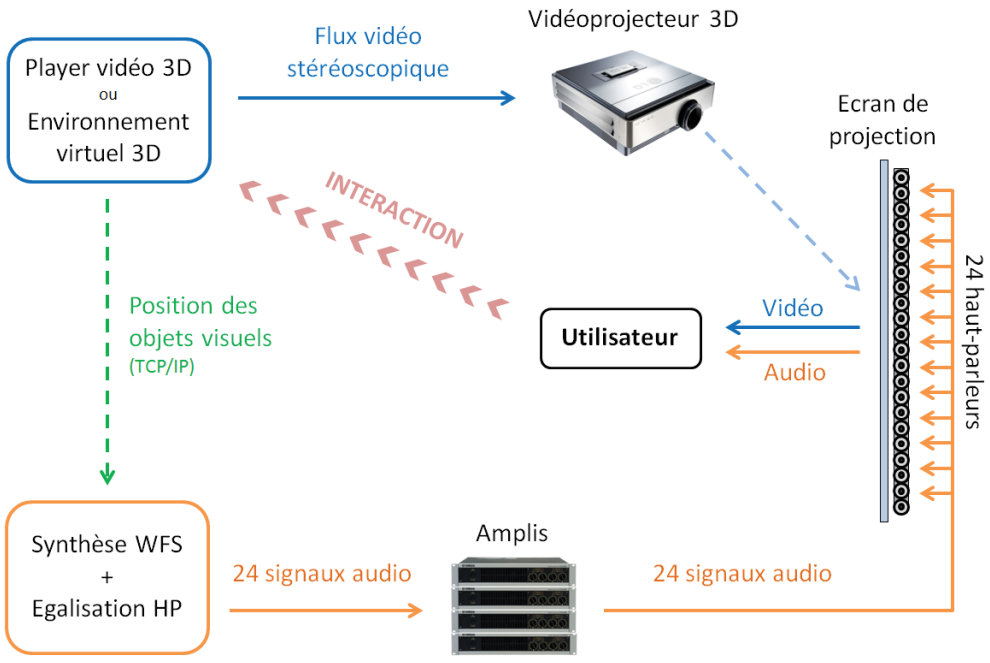


FIGURE 4.3 – *Architecture du système AV3D proposé, associant la vidéo stéréoscopique à un rendu sonore Wave Field Synthesis.*

4.4 Évaluation objective des performances du système WFS

Il est nécessaire d'évaluer objectivement les performances offertes par le système de restitution sonore proposé. Des mesures acoustiques sont donc réalisées dans le but de comparer les signaux audio issus de sources sonores réelles aux signaux issus de sources sonores virtuelles résultant de la synthèse par Wave Field Synthesis. Pour cela, un sinus glissant est diffusé par un haut-parleur réel ou par le système WFS. Le champ sonore résultant est enregistré au moyen d'une tête artificielle (Neumann KU-100) équipée de deux microphones omnidirectionnels (DPA 4053). Les signaux captés au niveau des oreilles droite et gauche sont ensuite convolués par un sinus glissant temporellement inversé afin d'en extraire les réponses impulsionnelles droite et gauche (HRIR ou *Head Related Impulse Response*). Cette mesure de HRIR est réalisée pour 10 positions de source, réparties sur une zone étendue : devant et derrière la barre WFS, ainsi qu'à droite et à gauche de l'axe central, afin de vérifier la symétrie du système ainsi disposé dans la salle. La figure 4.4 illustre les deux campagnes de mesure réalisées (sources sonores réelles et virtuelles). Les coordonnées des 10 positions de sources mesurées sont quant à elles présentées dans le tableau 4.1. L'origine du repère est définie au centre de la tête artificielle.

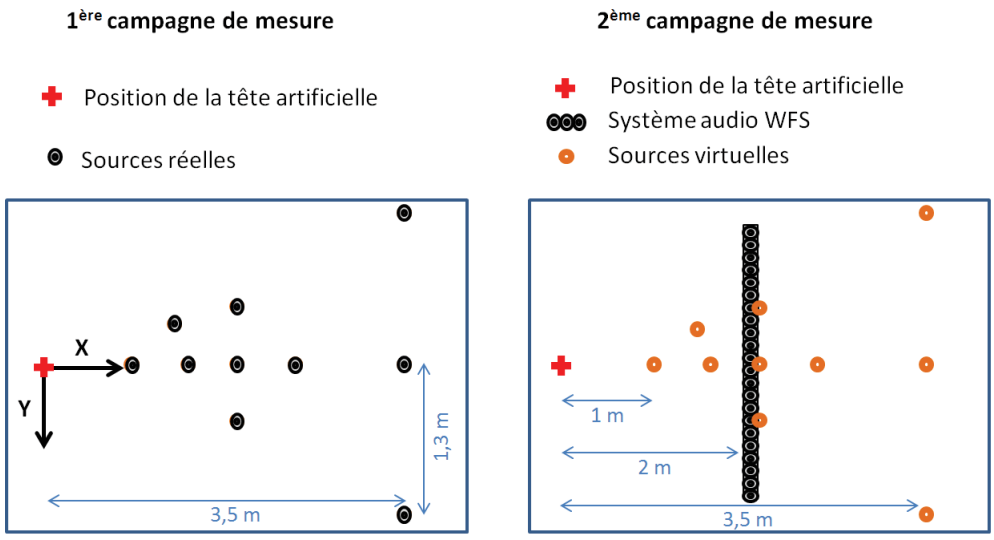


FIGURE 4.4 – *Campagnes de mesures acoustiques comparatives entre le champ sonore synthétisé par WFS et les sources réelles placées dans la salle d'expérimentation.*

Position	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
X (m)	1	1,4	1,6	2,1	2,1	2,1	2,5	3,5	3,5	3,5
Y (m)	0	-0,4	0	-0,5	0	0,5	0	-1,3	0	1,3

TABLE 4.1 – *Position des sources sonores réelles ou virtuelles utilisées pour la mesure des HRIR (Head Related Impulse Response).*

Une phase d'égalisation a été nécessaire afin d'uniformiser les réponses en fréquence entre les sources réelles et virtuelles. Pour cela, le spectre des haut-parleurs utilisés dans la première campagne de mesure (sources réelles) a été compensé pour correspondre aux sources sonores synthétisées par WFS. Le spectre cible appliqué à chaque source réelle a été déterminé par la réponse en fréquence d'une source sonore virtuelle placée à la position 5, et mesurée en un point (au centre de la tête artificielle).

Cette technique permet de réduire la complexité de la procédure d'égalisation. En effet, la méthode d'égalisation employée nécessite l'utilisation d'un seul filtre (appliqué au haut-parleur réel), alors que la compensation du spectre des sources virtuelles aurait nécessité l'utilisation de 24 filtres (appliqués aux haut-parleurs de la rampe de WFS). De plus, il aurait été difficile de déterminer chacun de ces filtres de manière individuelle puisque c'est en réalité la contribution combinée des 24 éléments qui aurait dû s'approcher au mieux du spectre cible.

À partir des HRIR droite et gauche mesurées, les informations d'ITD et d'ILD sont calculées. Le profil spectral propre à chaque position est également étudié pour les sources réelles comme pour les sources virtuelles. Ces données sont comparées dans la section suivante.

4.4.1 Comparaison d'ITD, ILD et ISSD entre sources réelles et virtuelles

ITD entre sources réelles et virtuelles

L'ITD ou *Interaural Time Difference* est estimée comme le retard relatif entre les HRIR gauche et droite. Parmi les différentes méthodes qui peuvent être utilisées pour évaluer ce retard, nous avons choisi l'estimation du maximum de la fonction d'intercorrélation des HRIR [Kistler and Wightman, 1992]. Ce choix repose notamment sur la robustesse de ce dernier en comparaison des estimateurs basés sur la phase [Nicol, 2010]. La JND (*Just Noticeable Difference*) de l'ITD définit la plus petite différence d'ITD perçue par le système auditif. Plusieurs études s'accordent à dire que la plus petite différence d'ITD détectable est de l'ordre de 10 μs [Klumpp and Eady, 1956, Hershkowitz and Durlach, 1969, Hafter and De Maio, 1975]. Cette valeur a été mesurée dans le cas d'un bruit large bande présenté en face d'un auditeur (ITD nulle). Il semblerait que ce seuil de détectabilité augmente avec l'azimut de la source virtuelle [Klumpp and Eady, 1956, Nicol, 2010]. De plus, la JND de l'ITD semble également dépendre du type de stimulus utilisé et de sa largeur de bande [Klumpp and Eady, 1956]. Compte tenu des valeurs reportées dans ces différentes études, il semblerait qu'une différence d'ITD inférieure à 10 μs ne soit pas détectable par l'oreille humaine.

Idéalement, l'écart d'ITD entre les sources réelles et les sources virtuelles doit être inférieur à cette valeur de JND ($\simeq$ 10 μs). Les valeurs d'ITD mesurées aux 10 positions de sources réelles et virtuelles sont présentées dans le tableau 4.2.

Les erreurs d'ITD présentées dans le tableau 4.2 sont inférieures à 5 μs pour toutes les positions mesurées. Ces résultats indiquent que la reproduction de l'ITD n'est pas dissociable, d'un point de vue perceptif, entre une source sonore réelle et une source sonore virtuelle synthétisée par notre système WFS.

Position	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
ITD source réelle (µs)	0,1	130,2	-0,1	111,2	-0,1	-109,9	-0,2	168,5	0	-167,3
ITD WFS (µs)	-3,8	129,7	4,1	115,9	-0,9	-113,6	2,3	170,7	4,5	-171,2
Erreur (µs)	-3,9	-0,5	4,2	4,7	-0,8	-3,7	2,5	2,2	4,5	-3,9

TABLE 4.2 – *ITD mesurées pour les sources sonores réelles et virtuelles aux 10 positions définies par le tableau 4.1.*

ILD entre sources réelles et virtuelles

L'ILD ou *Interaural Level Difference* est estimée comme le rapport en dB des spectres d'énergie des signaux gauche et droit sur la bande $[1 - 5\ kHz]$ [Larcher, 2001]. Différentes études ont montré que le seuil de détection d'un changement d'ILD est d'environ 1 dB [Mills, 1960, Yost and Dye Jr, 1988]. Ce seuil est minimal lorsque l'ILD de référence est de 0 dB, c'est-à-dire lorsque la source se trouve dans le plan médian. Ces études ont également montré une dépendance du JND d'ILD en fonction de la fréquence, avec des performances dégradées pour les fréquences autour de 1000 Hz. Dans le but d'étudier les performances offertes par notre système WFS, les erreurs d'ILD mesurées entre sources réelles et virtuelles sont présentées dans le tableau 4.3.

Position	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
ILD source réelle (dB)	0	4,8	0	4	0	-4	0	5,8	0	-5,7
ILD WFS (dB)	-0,4	3,6	0,9	4,1	0,1	-3,9	0,9	5,8	0,6	-5,9
Erreur (dB)	-0,4	-1,2	0,9	0,1	0,1	0,1	0,9	0	0,6	-0,2

TABLE 4.3 – *ILD mesurées pour les sources sonores réelles et virtuelles aux 10 positions définies par le tableau 4.1.*

Les erreurs d'ILD présentées dans le tableau 4.3 sont inférieures à 1 dB pour toutes les positions mesurées à l'exception de la position 2, pour laquelle une différence d'ILD de $1,2\ dB$ est mesurée entre la source réelle et la source virtuelle. Les valeurs d'ILD présentées dans ce tableau semblent indiquer que les performances de notre système WFS sont, dans l'ensemble, supérieures au seuil de détectabilité de changement d'ILD.

L'évaluation objective des indices de localisation (ITD et ILD) offerts par le système WFS proposé démontre sa capacité à restituer la latéralisation de sources sonores virtuelles. De plus, les résultats de ces mesures acoustiques indiquent que la spatialisation des sources dans le plan horizontal est effectuée avec une précision généralement supérieure aux seuils de discrimination du système perceptif humain.

ISSD entre sources réelles et virtuelles

Conjointement à la mesure acoustique des indices de localisation ITD et ILD, nous souhaitons quantifier, de manière objective, les performances de notre système WFS en termes de coloration spectrale. Il faut noter que les informations spectrales sont également porteuses des indices de localisation en élévation appelés Indices Spectraux (voir section 1.2.1).

L'idée est ici de mesurer la ressemblance entre le spectre d'une source virtuelle synthétisée par WFS et celui d'une source réelle. Les réponses en fréquence des sources sonores réelles et virtuelles sont mesurées à chaque oreille de la tête artificielle, et pour chacune des 10 positions de source (voir tableau 4.1). Les figures 4.5 et 4.6 illustrent respectivement les réponses en fréquence mesurées à la position 5 (où l'égalisation entre le système WFS et la source réelle cible a été réalisée) et à la position 8.

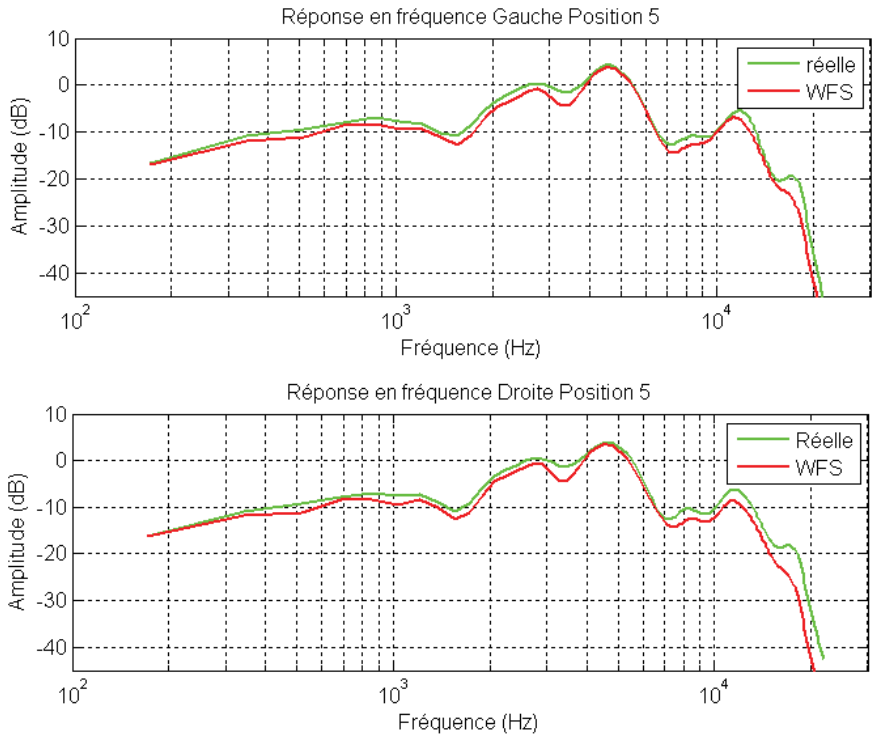


FIGURE 4.5 – *Réponses en fréquence mesurées à l'oreille gauche (en haut) et à l'oreille droite (en bas) de la tête artificielle, pour la source réelle (en vert) et la source virtuelle synthétisée par WFS (en rouge) à la position 5.*

Les réponses en fréquence présentées sur la figure 4.5 témoignent du bon fonctionnement de la procédure d'égalisation réalisée à la position 5. La figure 4.6 révèle la difficulté du système WFS à restituer les hautes fréquences (supérieures à environ 3 kHz), notamment dans le cas de la mesure réalisée à l'oreille gauche. Cette dernière correspond à l'oreille ipsilatérale pour la position 8. Cet effet est moins notable pour l'oreille controlatérale (droite) car les phénomènes de diffraction dus à la présence de la tête atténuent les composantes hautes fréquences de la source réelle.

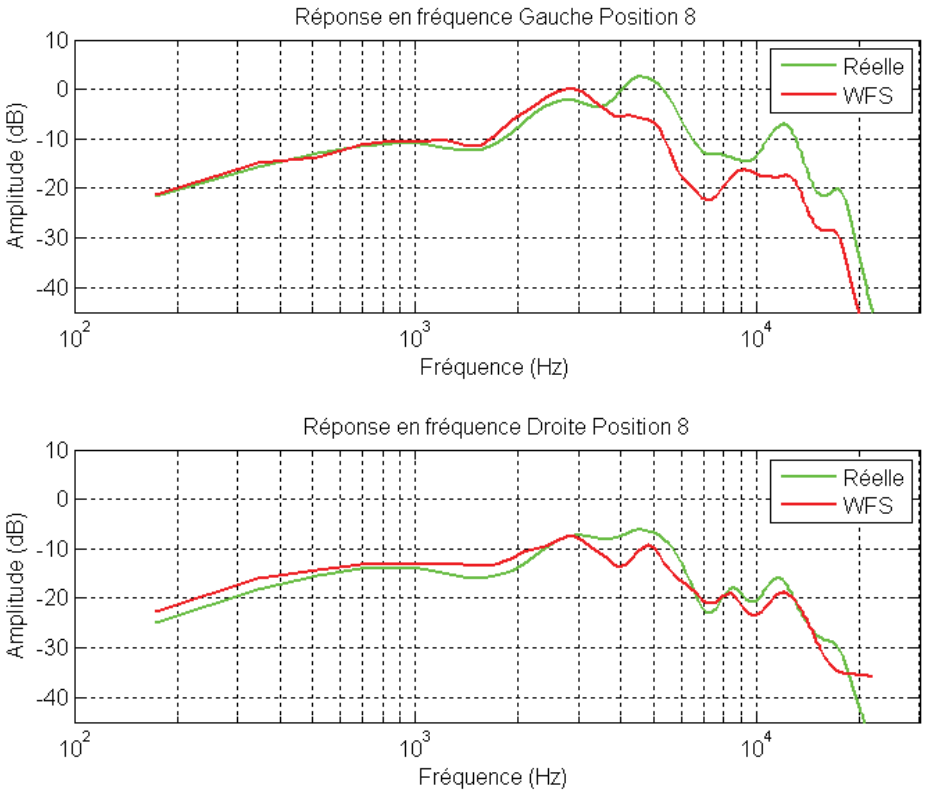


FIGURE 4.6 – *Réponses en fréquence des sources sonores pour la position 8, mesurées à l'oreille gauche (ipsilatérale) et à l'oreille droite (controlatérale) de la tête artificielle.*

Dans le but de mesurer la distance objective entre les spectres des sources réelles et virtuelles, le critère de l'*Inter-Subject Spectrum Difference* (ISSD) est calculé à partir des réponses en fréquence mesurées. Middlebrooks a initialement proposé l'ISSD pour quantifier la dissimilarité entre les Indices Spectraux (IS) de deux HRTF [Middlebrooks, 1999]. Dans notre cas, l'ISSD est appliquée pour évaluer la dissimilarité entre les spectres des signaux induits par une source sonore réelle et une source sonore virtuelle synthétisée par WFS. Ce critère est estimé comme la variance de la différence entre le spectre reproduit et le spectre cible considérés sur la bande $[4-13\ kHz]$ [Guillon, 2009] (bande fréquentielle porteuse des Indices Spectraux). Cependant, dans le cadre de notre étude, nous choisissons d'observer l'évolution du critère ISSD sur différentes bandes de fréquences : $[200\ Hz-1\ kHz]$, $[1-2\ kHz]$, $[2-3\ kHz]$, $[3-4\ kHz]$, $[4-7\ kHz]$ et $[7-13\ kHz]$. En effet, l'idée est de quantifier les colorations spectrales entre les sources virtuelles et les sources réelles sur une large zone de la bande audible, sans se limiter aux hautes fréquences. De plus, la fréquence d'*aliasing* d'environ 2 kHz de notre système WFS implique potentiellement une dégradation spectrale à partir de cette fréquence. Le tableau 4.4 répertorie les valeurs d'ISSD calculées sur les différentes bandes fréquentielles, et exprimées en dB^2 pour les 10 positions de sources définies par le tableau 4.1. Ces valeurs sont également représentées à l'aide de niveaux de couleurs sur la figure 4.7.

ISSD Oreille Droite

Position	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
[200-1kHz]	1,11	1,58	1,76	0,25	0,51	0,29	1,52	0,54	4,32	0,98
[1-2kHz]	0,46	0,80	0,18	0,19	0,15	0,19	0,64	0,62	1,21	0,50
[2-3kHz]	0,17	0,24	0,49	0,05	0,03	0,38	0,36	1,36	0,31	0,03
[3-4kHz]	3,80	2,64	2,77	12,76	0,88	16,09	6,59	3,46	4,29	3,08
[4-7kHz]	1,72	7,37	7,08	3,61	0,13	9,86	21,58	5,24	2,88	4,05
[7-13kHz]	11,35	1,18	20,80	4,88	0,32	22,54	4,67	3,32	7,35	8,68

ISSD Oreille Gauche

Position	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
[200-1kHz]	1,14	0,55	1,61	0,69	0,39	0,10	1,52	0,39	5,46	0,80
[1-2kHz]	0,32	0,22	0,25	0,25	0,20	0,24	0,66	0,53	1,33	0,46
[2-3kHz]	0,49	1,79	0,20	0,45	0,03	0,07	0,18	0,11	0,17	1,10
[3-4kHz]	4,26	0,98	1,56	16,39	0,50	14,72	7,38	4,81	4,38	3,67
[4-7kHz]	2,17	0,93	5,18	7,97	0,11	2,66	15,77	4,52	3,26	1,82
[7-13kHz]	2,50	11,84	12,21	17,60	0,64	9,39	1,94	9,74	3,13	3,26

TABLE 4.4 – *ISSD en dB^2 calculées sur différentes bandes de fréquences à partir des spectres mesurés aux oreilles droite et gauche de la tête artificielle.*

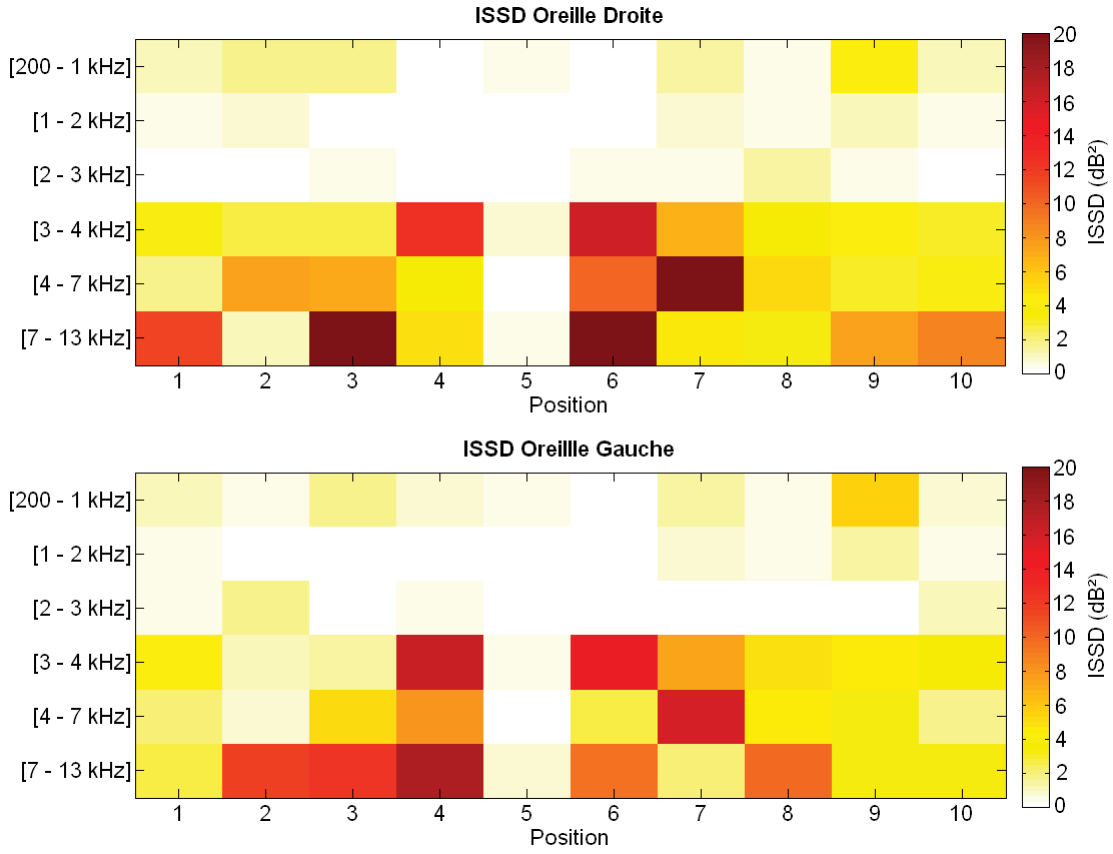


FIGURE 4.7 – *ISSD en dB^2 calculées sur les mesures des oreilles droite (haut) et gauche (bas), pour les 10 positions de sources et pour les différentes bandes de fréquences considérées.*

Les valeurs d'ISSD reportées dans le tableau 4.4 et sur la figure 4.7 semblent indiquer une dissimilarité spectrale plus importante entre les sources réelles et les sources virtuelles pour les fréquences supérieures à 3 kHz. Cette observation est valable pour toutes les positions de sources considérées à l'exception de la position 5, pour laquelle le critère d'ISSD est inférieur à 1 dB^2 sur l'ensemble des bandes considérées. Cette correspondance des spectres était attendue puisque la procédure d'égalisation a été réalisée à cette même position. La figure 4.8 synthétise les valeurs moyennes d'ISSD en fonction des bandes de fréquences considérées.

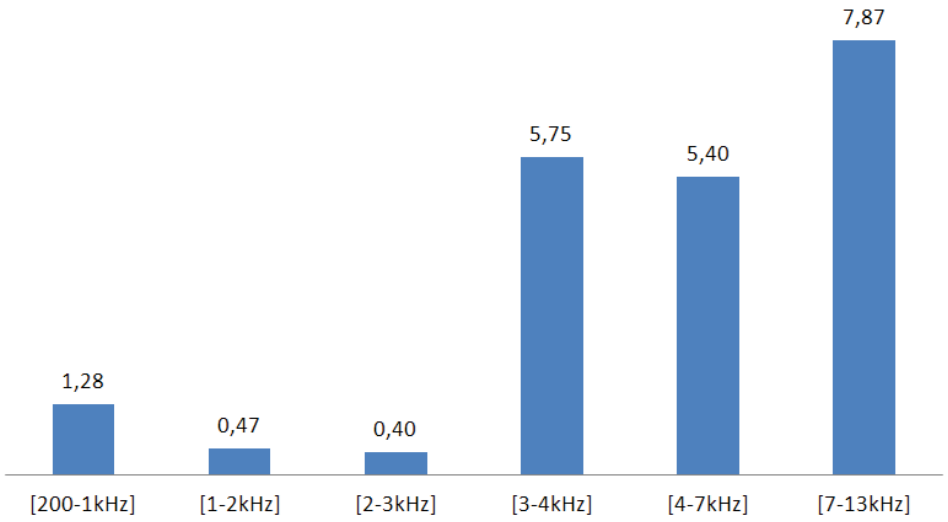


FIGURE 4.8 – *ISSD moyennes en dB^2 calculées sur différentes bandes de fréquences.*

La figure 4.8 illustre le phénomène de coloration apporté par le système WFS pour les bandes $[3-4\ kHz]$, $[4-7\ kHz]$ et $[7-13\ kHz]$. En effet, la distance objective entre le spectre des sources réelles et le spectre des sources virtuelles est beaucoup plus importante pour ces bandes. Au contraire, l'ISSD est minimale entre 200 Hz et 3 kHz, ce qui laisse présager des phénomènes de coloration moins marqués sur cette bande de fréquences. Notons cependant qu'il est difficile d'estimer dans quelle mesure ces dissimilarités sont détectables par l'oreille humaine car il n'existe pas, à notre connaissance, de seuil de détection (ou JND) clairement établi pour l'ISSD.

Néanmoins, il faut noter l'existence de certaines investigations menées sur la perception des différences spectrales liées aux HRTF. Guillon a, par exemple, comparé les performances de localisation dans le cas de HRTF individuelles ou non-individuelles (HRTF reconstruites) [Guillon, 2009]. Ces dernières impliquent des différences spectrales qui sont quantifiées par l'ISSD. Il est possible d'extraire de ces travaux les valeurs d'ISSD à partir desquelles les sujets ne perçoivent pas de différences entre les HRTF proposées (d'après la figure 6.27 et le tableau 6.2 de [Guillon, 2009]). Pour les cinq participants, ces valeurs d'ISSD sont inférieures à $2,5\ dB^2$ avec des valeurs comprises entre $2,6$ et $5\ dB^2$. Ces résultats semblent cohérents avec une récente étude de Rugeles et al. [Rugeles et al., 2014], dans laquelle différentes méthodes de lissage de HRTF sont évaluées au moyen d'un test d'écoute de type MUSHRA. Dans cette étude, les dégradations spectrales introduites par les méthodes de lissage sont quantifiées à l'aide

du critère logISSD. La conversion des valeurs de logISSD en valeurs d'ISSD indique qu'une dégradation spectrale correspondant à une ISSD de $4,07\ dB^2$ offre une qualité de spatialisation comparable à la référence (HRTF non-lissée). Bien que ces différentes études ne permettent pas de déterminer au sens strict le seuil de détectabilité d'ISSD, elles apportent néanmoins des éléments de réponse et permettent de fournir un ordre de grandeur quant aux valeurs d'ISSD tolérées suivant différentes tâches (localisation et évaluation de la qualité de spatialisation).

4.4.2 Discussions

Il a été montré que les indices de localisation binauraux (ITD et ILD) sont restitués de manière précise puisque les erreurs mesurées entre sources réelles et sources synthétisées par WFS sont inférieures aux valeurs de JND. Cette caractérisation objective (mesure des erreurs d'ITD et d'ILD) indique que le système WFS installé fournit des indices de localisation pertinents qui sont *a priori* exploitables par le système auditif.

La mesure des réponses en fréquence et le calcul de l'ISSD semblent témoigner de la bonne fidélité spectrale d'une source virtuelle par rapport à une source réelle dans une bande fréquentielle allant de 200 Hz à 3 kHz. Cette observation laisse penser que la synthèse du champ sonore par WFS est correctement réalisée au moins jusqu'à la fréquence d'*aliasing* (environ 2 kHz). Pour les fréquences supérieures à environ 3 kHz, le traitement WFS semble cependant introduire une coloration spectrale plus marquée. Cette coloration est potentiellement assimilable à une modification des Indices Spectraux (définis dans la section 1.2.1) qui jouent un rôle important dans la perception de la direction d'une source sonore et notamment de son élévation. Néanmoins, il est à noter que les expériences présentées dans la suite de ce travail ont pour problématique principale la perception de la distance d'objets sonores, et n'impliquent pas de tâches de localisation en élévation. De plus, aucune comparaison directe ne sera réalisée entre sources réelles et sources virtuelles.

Pour ces raisons, nous considérons que les performances générales du système WFS proposé sont suffisantes dans le cadre de notre travail. De plus, les mesures acoustiques réalisées témoignent de l'absence de problèmes dans la restitution sonore par Wave Field Synthesis (problèmes qui pourraient être liés au traitement WFS ou à une défaillance du matériel par exemple).

La suite de ce travail a pour but la validation de la restitution de la distance d'objets sonores et/ou visuels par le biais de tests subjectifs. Ces expériences sont menées grâce au système de restitution audio-visuelle proposé (AV3D) et sont décrites dans le chapitre 5.

Chapitre 5

Perception d'objets virtuels placés suivant la distance grâce au système AV3D

5.1 Introduction

L'expérimentation réalisée dans le chapitre 3 a permis de mettre en évidence l'impact de la cohérence entre le son et l'image sur la sensation d'immersion audio-visuelle perçue par les spectateurs. Afin d'améliorer la qualité d'expérience liée à la visualisation de contenus vidéo 3D, nous souhaitons proposer une restitution audio-visuelle spatialement cohérente, en considérant notamment la nouvelle dimension introduite par la vidéo relief : la distance. L'idée consiste à introduire des objets sonores suivant la distance dans le but d'accompagner au mieux le relief des images 3D.

Pour cela, nous avons tout d'abord proposé dans le chapitre 4 un système de restitution audio-visuelle 3D associant la technologie Wave Field Synthesis au rendu visuel stéréoscopique (système AV3D). Les performances offertes par le système WFS ont été évaluées de manière objective. Les indices de localisation binauraux (ITD, ILD) associés aux sources sonores synthétisées semblent pertinents puisque les valeurs d'erreurs mesurées sont inférieures aux seuils perceptibles.

Cette section a pour objectif d'évaluer, de manière subjective, la capacité du système de restitution proposé (AV3D) à restituer des informations sonores et/ou visuelles suivant la distance. Dans ce but, un protocole expérimental doit dans un premier temps être défini. Une première expérience est réalisée sur la perception de la distance de sources sonores réelles (haut-parleurs) afin de s'assurer de la pertinence de la méthode de recueil des jugements retenue. Le protocole proposé est ensuite utilisé pour évaluer la perception de la distance d'objets virtuels, dans le cas de présentations unimodales ou bimodales. Trois expérimentations sont alors réalisées.

5.2 Validation du protocole expérimental

5.2.1 Méthodes d'estimation de la distance

Il existe de nombreuses méthodes d'estimation de la distance. Quatre types de protocoles peuvent être répertoriés [Klein et al., 2009] : l'estimation verbale, les actions guidées, les actions imaginées, les correspondances perceptuelles.

Les méthodes basées sur l'**estimation verbale** (également appelées méthodes de verbalisation) se font à l'aide d'une échelle de mesure utilisant une unité familière (le mètre, par exemple). Cette catégorie inclut également les méthodes de "report direct", dans lesquelles les jugements sont effectués sans utiliser d'échelle [Gogel and Tietz, 1973, Zahorik, 2002a].

Dans les protocoles basés sur les **actions guidées**, les participants doivent effectuer une action pour estimer la distance de l'objet. Il peut s'agir de marcher directement jusqu'à la cible [Loomis et al., 1998] (méthode de "marche aveugle"), ou de suivre un chemin indirect [Rébillat et al., 2012] (méthode de "triangulation").

Au contraire des méthodes impliquant des actions guidées, les protocoles basés sur les **actions imaginées** permettent aux participants d'estimer la distance d'objets sans avoir à se déplacer. Les actions imaginées peuvent être de différente nature. Les participants peuvent, par exemple, estimer le temps de marche ou le nombre de pas à effectuer pour atteindre une cible.

Dans les protocoles basés sur la **correspondance perceptuelle**, les sujets doivent estimer la distance d'un objet cible en effectuant une comparaison avec un objet témoin. Ces protocoles regroupent les "méthodes d'ajustement" [Wittek et al., 2004] et les méthodes utilisant le placement d'un objet à mi-distance (*perceptual bisection*) [Bodenheimer et al., 2007, Lappin et al., 2006].

Il est important de noter que toutes les méthodes d'estimation de la distance ont des avantages et des inconvénients. Les méthodes d'actions guidées nécessitent l'utilisation d'autres modalités et peuvent donc apporter des biais supplémentaires liés à la motricité par exemple. Les résultats issus d'actions imaginées nécessitent quant à eux d'être convertis en distance. Les méthodes de report verbal sont plus simples à mettre en œuvre et ne nécessitent pas de conversion, mais aboutissent en général à des résultats plus variables. Malgré la variabilité importante des jugements issus d'estimations verbales, Loomis et al. ont montré qu'ils sont comparables à ceux obtenus par des méthodes de marche directe ou indirecte [Loomis et al., 1998].

5.2.2 Modélisation de la perception de la distance d'objets sonores et visuels

Il y a quelques années, Zahorik a entrepris un travail bibliographique [Zahorik et al., 2005] regroupant 84 articles sur la perception de la distance de sources sonores réelles. Les conditions expérimentales de ces études sont très variées en termes de stimuli, d'environnements expérimentaux, ou encore de méthodes d'estimation. Cependant, la plupart des résultats indiquent que les participants ont tendance à surestimer la distance des sources sonores proches (placées à moins de $1,5\ m$ environ), alors que la distance des sources placées plus loin est sous-estimée. D'après Zahorik, la loi de puissance de Stevens [Stevens, 1957] est une bonne approximation de

la relation entre la distance perçue et la distance physique des sources sonores. Cette loi s'écrit sous la forme $d_p = k.{d_r}^a$, où d_p est la distance perçue, d_r est la distance physique, l'exposant a est le facteur de compression de la fonction, et k est une constante. Zahorik analyse les résultats de son corpus d'articles et trouve que la valeur moyenne de k est de 1,32, et que l'exposant a est toujours inférieur à 1 ($\overline{a} = 0,54$), ce qui traduit la compression de la perception de la distance.

Plusieurs études ont récemment été menées sur la perception de la distance des sources sonores dans les environnements virtuels. Ces études mettent en œuvre de nombreuses technologies audio 3D comme la technologie binaurale [Côté et al., 2011, Zahorik, 2002a, Anderson and Zahorik, 2011], ou encore les méthodes basées sur l'holophonie dans le cas de sources statiques [Kearney et al., 2012, Komiyama et al., 1991], et en mouvement [Rébillat et al., 2012, Corteel, 2004]. Ces études montrent que les performances de localisation auditive en distance sont comparables entre les environnements réels et virtuels. Les distances sont en règle générale sous-estimées pour les sources sonores placées à plus de 2 m, et les résultats sont en accord avec le modèle de compression évoqué par Zahorik.

La littérature portant sur la localisation en distance d'objets visuels dans un environnement réel ou virtuel est relativement fournie [Da Silva, 1985, Klein et al., 2009]. La distance d'un objet visuel est relativement bien perçue dans les environnements réels, mais est sous-estimée dans les environnements virtuels (casques de réalité virtuelle, CAVE, etc.) [Cutting and Vishton, 1995, Renner et al., 2013, Loomis and Knapp, 2003]. La perception de la distance visuelle est généralement modélisée par la loi de Stevens comme dans le cas de l'audition [Teghtsoonian and Teghtsoonian, 1969, Wiest and Bell, 1985]. Le modèle de compression qui relie la distance restituée à la distance perçue s'écrit ainsi $d_p = k.{d_r}^a$.

5.2.3 Protocole expérimental proposé

La méthode d'estimation de la distance doit, pour les besoins de nos expériences, être applicable dans le cas d'objets virtuels. Cette contrainte rend par exemple impossible l'utilisation de la méthode de marche aveugle [1]. Nous choisissons d'utiliser une méthode d'estimation verbale, la méthode du report direct. Le recueil des jugements est simplifié puisque les participants doivent indiquer directement au clavier la distance des objets qui leur sont présentés. Les distances doivent être estimées en mètres, avec une précision à la décimale. Les participants sont statiques et n'ont pas besoin de se déplacer comme dans les méthodes basées sur les actions guidées.

Afin de limiter la variabilité intra-individuelle des jugements récoltés par la méthode de report direct, chaque condition de test est répétée 5 fois (toutes les conditions sont présentées dans un ordre aléatoire). Ensuite, il est important de garder à l'esprit que les jugements de distance sont relativement variables suivant les individus (voir section 1.2.3). Pour prendre en compte cette variabilité inter-individuelle, un panel relativement conséquent est utilisé (24 participants).

1. Il est impossible pour les participants d'atteindre un objet virtuel placé derrière le dispositif. En effet, ce dernier représente un obstacle physique positionné entre les sujets et la cible virtuelle.

Nous proposons de tester différents stimuli car, comme mentionné dans les sections 1.2.2 et 1.2.3, certaines études laissent penser que la nature du stimulus peut avoir une influence sur la perception de la distance, notamment pour les objets sonores. Ainsi, suivant les modalités étudiées dans les expériences, trois stimuli audio et/ou visuels sont présentés :

— "Avatar/Voix" (personnage virtuel accompagné d'un signal de parole),
— "Capsule/Bruit" (cylindre flottant dans l'air émettant des salves de bruit blanc),
— "Tel/Sonnerie" (téléphone gris flottant dans l'air et sonnerie de téléphone).

Ces stimuli ont été choisis pour leurs différences marquées en termes de taille, de contenu spectral, mais aussi d'un point de vue cognitif. Un personnage qui parle, par exemple, est supposé plus familier qu'un cylindre émettant des salves de bruit blanc. En effet, la taille d'une personne et le niveau sonore qu'elle peut générer en situation de conversation nous sont *a priori* connus. La relation sémantique existant entre le stimulus sonore et le stimulus visuel est également variable suivant le couple de stimuli. Il n'y a en effet aucun lien naturel entre le cylindre et le bruit qui l'accompagne contrairement aux liens existants entre une personne qui parle et un son de parole, ou encore entre un téléphone et une sonnerie de téléphone. La figure 5.1 représente les stimuli visuels utilisés et propose une représentation temps-fréquence (spectrogramme) des stimuli sonores associés.

Les stimuli sont présentés dans des sessions différentes, dont l'ordre est alterné aléatoirement en s'assurant que chaque combinaison possible soit évaluée par le même nombre de participants.

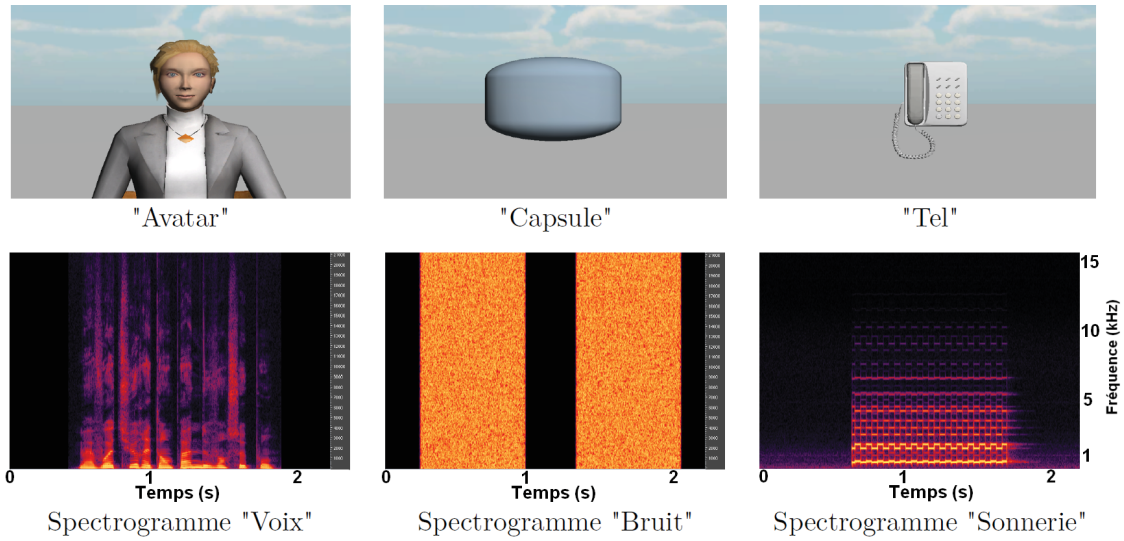


FIGURE 5.1 – *Illustration des stimuli visuels (en haut) utilisés dans les expériences impliquant la modalité visuelle seule ou une présentation bimodale, et les spectrogrammes (en bas) associés aux stimuli sonores.*

Préalablement à la phase de notation, une phase de familiarisation aux stimuli est proposée lors de laquelle tous les objets virtuels sont présentés à toutes les positions dans un ordre aléatoire. Une phase de familiarisation à la tâche est ensuite réalisée : le participant est mis en situation de test. Il peut ainsi prendre en main l'interface de test et s'entraîner à estimer la distance des objets virtuels.

5.2.4 Perception de la distance d'une source sonore réelle

Une première expérience subjective est réalisée dans le but de s'assurer de la pertinence et de la robustesse du protocole d'estimation de la distance retenu.

Environnement expérimental

Dans cette expérimentation, la tâche des participants consiste à évaluer la distance de huit sources sonores réelles (haut-parleurs) disposées en face d'eux à 1, 1,5, 2, 2,5, 3, 3,5, 4 et 5 m. La figure 5.2 permet de se rendre compte du placement des haut-parleurs. Les transducteurs sont légèrement décalés en élévation et en azimut afin de réduire la diffraction des ondes sonores par le corps des haut-parleurs. Ces décalages sont limités à une fenêtre spatiale de $\pm 1,5°$ en azimut et $\pm 2,5°$ en élévation, ce qui reste inférieur au flou de localisation du système auditif (voir sections 1.2.2 et 1.4.3). Ces décalages sont, en principe, imperceptibles par les participants. L'approximation suivant laquelle les haut-parleurs sont placés en face des participants peut donc être faite.

FIGURE 5.2 – *Haut-parleurs placés en face des participants, à une distance allant de 1 à 5 mètres.*

Les réponses en fréquence des huit haut-parleurs sont égalisées par rapport à un haut-parleur "cible". Pour ce faire, le haut-parleur "cible" est successivement placé à chacune des huit positions testées. Les huit spectres "cibles" sont alors mesurés au point d'écoute. Lors de l'étape de synthèse des sources virtuelles, ce sont ces spectres cibles qui sont visés, en considérant pour la position n le spectre induit par le haut-parleur situé à la position n. Cette méthode d'égalisation permet, lors de la phase de test, de recréer artificiellement la situation d'un haut-parleur qui est physiquement déplacé aux différentes positions testées. De plus, la procédure d'égalisation tient compte de l'onde directe uniquement, ce qui permet de respecter le comportement acoustique naturel de la salle en fonction de la position des haut-parleurs. Les haut-parleurs sont également calibrés pour émettre un niveau sonore de 68 $dB(A)$ (mesuré à 1 mètre).

L'expérience se déroule dans la salle décrite dans le chapitre 3 (superficie d'environ 20 m^2, niveau de bruit de fond inférieur à 30 $dB(A)$, $Tr_{60} \simeq 350\ ms$). Un drap noir (transparent d'un point de vue acoustique) est installé devant les haut-parleurs de manière à dissimuler la position de ces derniers aux participants. Ainsi, aucun indice visuel ne peut biaiser le jugement de la distance des sources sonores.

Stimuli

Cette première expérience ne portant que sur la modalité auditive, les trois stimuli utilisés sont "Voix", "Bruit" et "Sonnerie" (voir figure 5.1). Ces stimuli sont présentés dans des sessions de test séparées. Au sein de chaque session, le stimulus sonore est diffusé de manière continue jusqu'au jugement de distance du participant.

Panel

Le test est réalisé sur un panel de 24 participants rémunérés dont la moyenne d'âge est de 31 ans (11 femmes et 13 hommes). Ces personnes ont toutes l'habitude d'effectuer des tests d'écoute, mais aucune n'avait réalisé de test sur la perception de la distance d'un objet sonore auparavant.

Résultats

La durée totale de test est en moyenne de 30 minutes (pauses comprises). Les jugements de distance sont soumis à une analyse de la variance (ou ANOVA pour *ANalysis Of VAriance*). La figure 5.3 représente les moyennes et intervalles de confiance à 95 % des distances estimées par 22 testeurs pour les trois stimuli sonores étudiés. En effet, deux participants ont été exclus de cette analyse car leurs jugements dépassaient de plus de 5 fois la variabilité inter-individuelle du panel [2]. Notons que l'observation de jugements aberrants était en partie prévisible puisque le protocole expérimental utilisé laisse une liberté totale aux participants quant à la gamme de distances évaluées. En effet, aucune indication n'est fournie sur la distance minimale ou maximale des haut-parleurs dans le but de recueillir les jugements spontanés des participants.

La ligne grise sur la figure 5.3 représente la relation identité entre la distance réelle et la distance perçue des sources sonores. Les résultats de cette expérience indiquent que les participants ont tendance à sous-estimer la distance des haut-parleurs, en particulier pour les distances supérieures à environ 2 mètres. Ces résultats semblent en accord avec la littérature sur le sujet [Zahorik et al., 2005].

Des différences peuvent être observées sur la figure 5.3 suivant les stimuli testés. En effet, la sonnerie de téléphone, par exemple, semble globalement perçue comme étant légèrement plus distante que le bruit. Cependant, cette tendance n'est pas assez marquée pour être statistiquement significative puisque l'analyse ANOVA ne révèle pas d'effet du stimulus ($F(2,42)=2.96$, $p=0.06$).

2. Jugements estimés aberrants au regard des résultats pour le haut-parleur le plus proche (situé à 1 m), pour lequel la variabilité inter-individuelle est la plus faible.

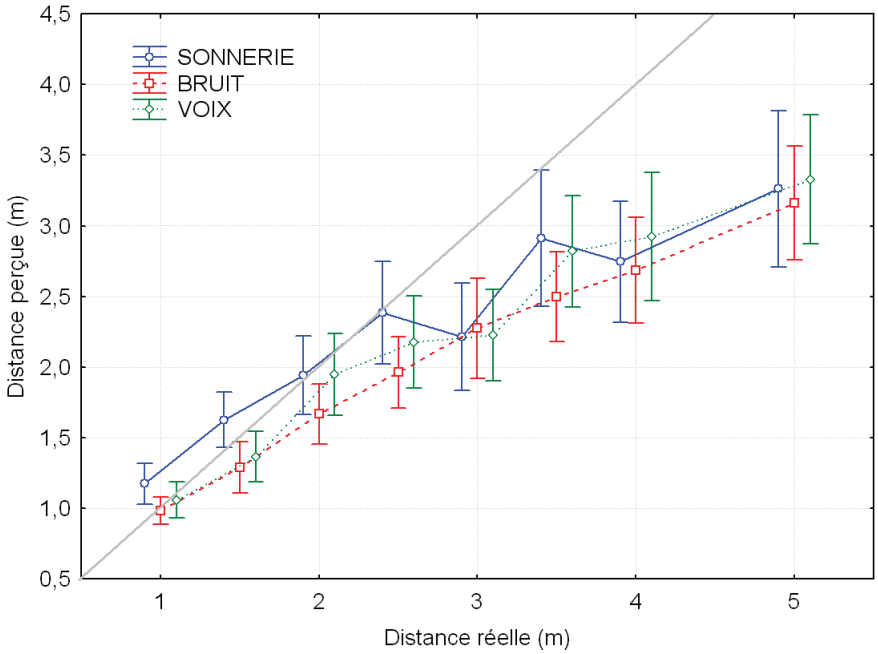


FIGURE 5.3 – *Distances perçues des sources sonores réelles et intervalles de confiance à 95 % pour les stimuli "Sonnerie" (en bleu), "Bruit" (en rouge) et "Voix" (en vert).*

Nous pouvons également noter que les jugements sont plus précis, à la fois en termes d'erreur de localisation et de variabilité des réponses, pour les sources sonores les plus proches. Cette augmentation de la variabilité des réponses en fonction de la distance réelle des sources est illustrée sur la figure 5.4. Il faut noter que les intervalles de confiance à 95 % représentent entre 20 et 30 % de la distance réelle du haut-parleur pour tous les stimuli. Ces valeurs sont tout à fait raisonnables en comparaison des valeurs avancées par Zahorik (entre 20 et 60 % suivant les participants) [Zahorik, 2002b].

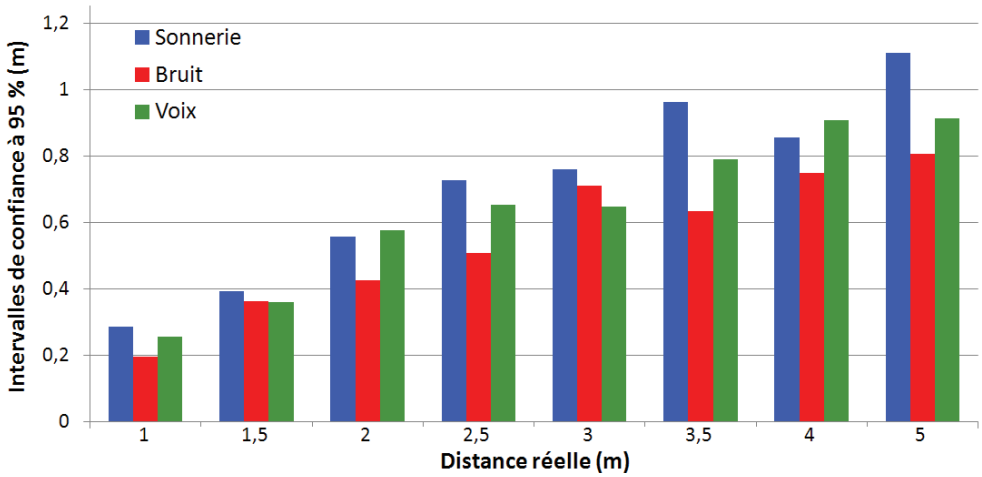


FIGURE 5.4 – *Intervalles de confiance à 95 % exprimés en mètres, tirés de la figure 5.3.*

Il est maintenant intéressant de comparer ces résultats expérimentaux au modèle de perception de la distance régulièrement évoqué dans la littérature (section 5.2.2). L'idée est ici de savoir si les jugements collectés grâce au protocole proposé sont en accord avec la loi de puissance utilisée par Zahorik pour modéliser la distance perçue de sources sonores réelles. Le tableau 5.1 présente les valeurs des coefficients k et a du modèle, ainsi que les coefficients de détermination (R^2) associés pour les trois stimuli sonores. Le coefficient R^2 est utilisé pour évaluer la fiabilité d'un modèle de prédiction par rapport à des données expérimentales. Ses valeurs varient de 0 à 1. Une valeur de 0 indique l'absence totale de corrélation entre les données et le modèle alors qu'une valeur de 1 indique une concordance parfaite.

Stimulus	*k*	*a*	*R²*
Sonnerie	1.237	0.62	0.958
Bruit	0.986	0.74	0.997
Voix	1.075	0.73	0.979

TABLE 5.1 – *Coefficients des fonctions de compression et coefficients de détermination des trois stimuli sonores dans le cas de sources sonores réelles.*

Les valeurs de la constante k (de l'ordre de 1) et de l'exposant a semblent en accord avec les résultats avancés par Zahorik. De plus, les valeurs élevées du coefficient de détermination ($R^2 > 0,95$) indiquent que la loi de compression de Stevens permet une bonne approximation des valeurs expérimentales.

Pour conclure, les résultats de cette expérience sont cohérents avec les travaux similaires trouvés dans la littérature. Le protocole d'estimation de la distance proposé semble donc valide puisqu'il permet de recueillir des jugements de distance fiables et d'observer des tendances générales comparables aux travaux précédemment menés sur le sujet. Aussi, le protocole expérimental décrit dans la section 5.2.3 est appliqué, dans la suite de ce chapitre, pour évaluer la distance d'objets virtuels unimodaux (sonores ou visuels) et bimodaux (audio-visuels).

5.3 Présentation unimodale

Deux expériences sont réalisées sur la perception de la distance en condition unimodale. L'objectif est de savoir si le système AV3D proposé dans le chapitre 4 est capable d'offrir une restitution de la distance pour des objets sonores d'un côté, et visuels de l'autre. Le tableau 5.2 répertorie les distances testées en fonction de la modalité étudiée.

Exp.	*Modalité*	*Distances des objets (m)*
1	Audition	1, 1.5, 2, 2.5, 3, 3.5, 4, 5, 7, 10
2	Vision	2, 2.5, 3, 3.5, 4, 5, 6, 7

TABLE 5.2 – *Distances simulées dans les deux expériences portant sur la perception de la distance d'objets unimodaux (sonores ou visuels).*

Ce tableau montre que l'étendue des distances diffère suivant la modalité. Ce choix est motivé, d'une part, par les contraintes imposées pour une visualisation confortable des objets 3D (voir section 4.3.2), et d'autre part, par notre volonté de pouvoir présenter des objets sonores placés devant et derrière chaque objet visuel. Cette problématique liée à la présentation d'objets audio-visuels spatialement incohérents sera traitée dans le chapitre suivant. Néanmoins, nous proposons d'étudier dès à présent la perception de la distance d'objets sonores sur cette gamme de distance plus étendue (entre 1 et 10 m). Notons également que la répartition des distances simulées est irrégulière dans les deux expériences. Cette répartition permet d'offrir une description de la perception de la distance à proximité du dispositif de restitution AV3D, tout en tenant compte de l'augmentation du flou de localisation avec la distance (voir figure 5.4).

Les deux expériences sont menées suivant le protocole décrit dans la section 5.2.3. Elles sont réalisées à des dates différentes sur des panels de 24 participants rémunérés.

5.3.1 Perception de la distance d'un objet sonore synthétisé par WFS

Dans cette première expérience, les participants doivent estimer la distance de sources sonores synthétisées par Wave Field Synthesis. La position d'écoute est à 3 m de la rampe de haut-parleurs et les sources sonores à localiser sont simulées en face des participants, entre 1 et 10 m. Ainsi, les objets sonores sont synthétisés devant et derrière le système de restitution comme l'illustre la figure 5.5. Un drap acoustiquement transparent est installé devant les participants de manière à dissimuler la rampe de haut-parleurs. Le niveau sonore des sources sonores virtuelles restituées est calibré. Il correspond au niveau sonore d'un monopôle émettant 73 $dB(A)$ à 1 m, et respecte une loi d'atténuation arbitraire de 6 dB par doublement de distance [3].

3. La loi d'atténuation mesurée dans la salle d'expérimentation est de l'ordre de $4,5$ dB par doublement de distance. Cependant, cette expérience n'a pas pour vocation de recréer un environnement sonore parfaitement identique à celui de la salle d'expérimentation, mais plutôt d'explorer la capacité du système WFS à fournir des informations relatives à la distance d'objets sonores. Une décroissance de 6 dB par doublement de distance est utilisée car elle garantit une meilleure lisibilité des informations de distance.

Les trois stimuli sonores "Voix", "Bruit" et "Sonnerie" sont présentés à un panel de 24 participants (13 femmes et 11 hommes), dont la moyenne d'âge est d'environ 30 ans.

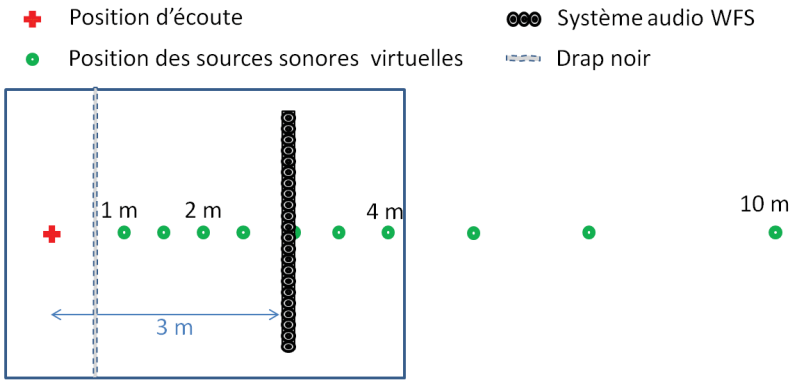


FIGURE 5.5 – *Dispositif Wave Field Synthesis et position des sources sonores à localiser.*

Résultats

Une analyse de variance (ANOVA) est effectuée sur les jugements de tous les participants, en considérant deux facteurs, le stimulus sonore utilisé et la distance synthétisée. La figure 5.6 représente les distances moyennes perçues des sources sonores virtuelles et les intervalles de confiance à 95 % associés.

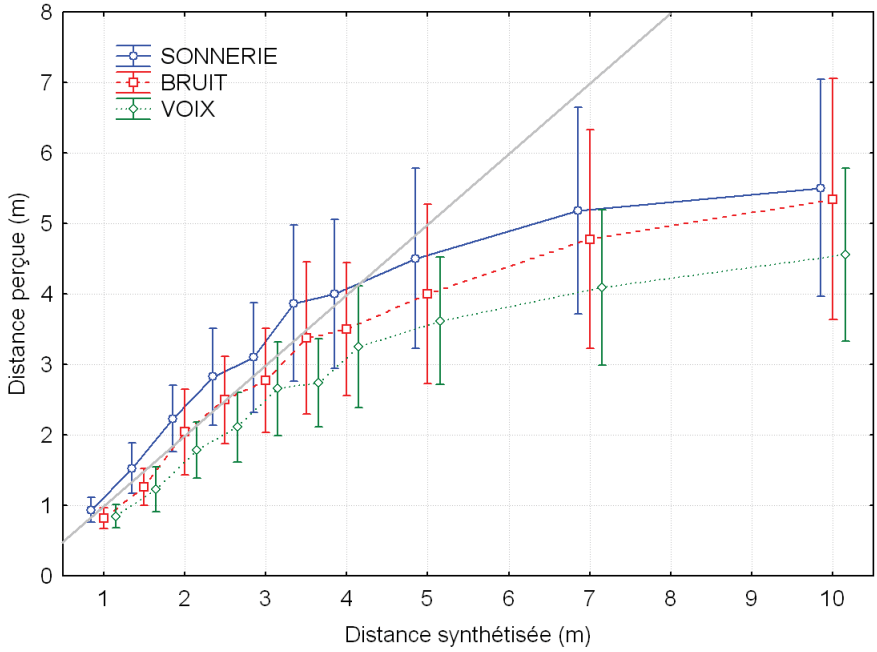


FIGURE 5.6 – *Distances perçues et intervalles de confiance à 95 % des sources sonores synthétisées par WFS pour les stimuli "Sonnerie" (en bleu), "Bruit" (en rouge) et "Voix" (en vert).*

Pour commencer, les distances perçues sont en règle générale sous-estimées pour tous les stimuli à partir de 4 m. Ensuite, les intervalles de confiance indiquent que les jugements des participants sont plus précis pour les sources sonores les plus proches. Cette figure montre également que les différences perçues entre les stimuli s'amplifient avec la distance. Le stimulus "Voix" est généralement perçu comme étant plus proche que les stimuli "Bruit" et "Sonnerie". De plus, la variabilité des résultats pour le signal de parole est moins importante, en particulier pour les distances de 5, 7, et 10 mètres. Cet effet du stimulus est confirmé par l'analyse ANOVA ($F(2,46)=6.13$, $p<0.005$). Un test *post-hoc* de type Bonferroni révèle que cet effet est uniquement présent entre les stimuli "Voix" et "Sonnerie". Il a été montré que la perception de la distance d'un signal de parole peut être influencée par l'intonation du locuteur [Gardner, 1969, Brungart and Scott, 2001]. Les conclusions de ces travaux montrent notamment qu'une voix "parlée" est naturellement perçue comme étant plus proche qu'une voix "criée". Dans notre cas, le stimulus de parole utilisé est comparable à une voix "parlée". De ce fait, pour les distances supérieures à quelques mètres, il est possible qu'une sous-estimation de la distance vienne s'ajouter au phénomène de compression décrit précédemment, du fait de l'intonation. En effet, dans un environnement naturel, une personne placée à plusieurs mètres aurait plutôt tendance à crier pour se faire entendre. Concernant la variabilité des estimations, Coleman [Coleman, 1962] a montré que l'utilisation de stimuli sonores familiers peut potentiellement augmenter la précision des jugements, ce qui serait en accord avec les résultats présentés ici.

Plusieurs travaux ont déjà abordé le cas de la perception de la distance de sources sonores virtuelles avec des technologies de restitution variées (voir section 5.2.2). Afin d'évaluer les performances de notre système WFS par rapport à ces travaux (en termes de distance restituée perçue), le tableau 5.3 propose les valeurs des coefficients k et a du modèle de Stevens, ainsi que les coefficients de détermination (R^2) associés à ces études. Les références bibliographiques annotées (*) indiquent que les données expérimentales initiales ont été traitées afin de calculer les coefficients a, k et R^2.

Il apparaît, à la lecture du tableau 5.3, que l'exposant a est relativement élevé dans le cas de notre étude, avec des valeurs comprises entre 0,75 et 0,82 suivant les stimuli sonores. Ces valeurs indiquent que la compression de la distance perçue est moins marquée en comparaison des études précédentes utilisant la WFS comme système de restitution sonore. En d'autres termes, notre système offre une dynamique de distances perçues plus importante que dans les études précédentes. Notons cependant que les valeurs de a sont, comme dans toutes les études issues de la littérature, inférieures à 1. Elles traduisent donc le phénomène de compression de la distance perçue. Les coefficients k sont également cohérents avec les travaux cités précédemment. Les coefficients de détermination R^2 sont supérieurs à 0,9 pour tous les stimuli testés dans notre étude. Ce résultat indique que la distance d'objets sonores virtuels diffusés par notre système WFS peut être modélisée, comme dans le cas de sources sonores réelles, par la loi de Stevens.

Source bibliographique	*Restitution sonore*	*Environnement*	*Méthode de report*	*Stimulus*	*Gamme distances*	*a*	*k*	*R^2*
[Anderson, 2011]	Binaural	Salle de concert virtuelle	Report direct	Bruit blanc	0.3 - 9.75	0.66	2.17	0.66
[Rébillat et al., 2012]	WFS	Salle d'écoute	Triangulation	Bruit blanc filtré passe-bas et modulé (15 Hz)	1.5 - 5	0.33	1.72	0.98
[Côté et al., 2011]*	Binaural	Salle virtuelle Tr_{60} = 370ms	Report direct	Parole	2.0 - 20	0.81	0.77	0.98
	-	Salle virtuelle Tr_{60} = 860ms	-	-	-	0.62	1.51	0.95
[Wittek et al., 2004]*	WFS	Salle anéchoïque	Ajustement	Salves de bruit rose	0.25 - 1.90	0.18	1.00	0.99
[Kearney et al., 2012]*	HOA	Salle d'écoute	Marche directe	Parole	2.0 - 8.0	0.73	1.16	0.99
Moulin et al. (étude présente)	WFS	Salle d'écoute T60 = 350ms	Report direct	Parole	1.0 - 10	0.75	1	0.95
	-	-	-	Salves de bruit blanc	-	0.82	1.03	0.94
	-	-	-	Sonnerie de téléphone	-	0.78	1.21	0.92

TABLE 5.3 – *Conditions expérimentales et résultats issus de différentes sources bibliographiques sur la perception de la distance de sources sonores virtuelles. Les coefficients a et k des fonctions de compression ainsi que les coefficients de détermination R^2 déterminés dans la présente étude sont également présentés.*

Pour conclure, les résultats de cette expérience semblent indiquer que le système Wave Field Synthesis proposé est capable de synthétiser des sources sonores perçues à différentes distances, devant et derrière le dispositif de restitution placé à 3 m. De plus, la perception de la distance de ces objets est en accord avec les études précédentes impliquant la localisation de sources sonores virtuelles, mais aussi réelles. Cette observation témoigne de la pertinence des indices sonores restitués par WFS d'un point de vue perceptif.

5.3.2 Perception de la distance d'un objet visuel 3D

Dans cette seconde expérience, les participants doivent estimer la distance d'objets visuels présentés grâce à un vidéoprojecteur 3D. La distance de visualisation est fixée à 3 m de l'écran de projection et les objets sont simulés entre 2 et 7 m, c'est-à-dire de part et d'autre de l'écran (voir tableau 5.2). Cette gamme de distances a été déterminée sur la base des conditions de visualisation (taille de l'écran, distance de visualisation, etc.), dans le but de limiter les disparités binoculaires à des valeurs "confortables" (voir section 4.2.1).

Les trois stimuli visuels "Avatar", "Capsule" et "Tel" sont présentés à un panel de 24 participants (12 femmes et 12 hommes), dont la moyenne d'âge est d'environ 31 ans.

Résultats

Une analyse de la variance (ANOVA) est réalisée en considérant deux facteurs : le stimulus visuel et la distance simulée. Les résultats de cette analyse sont illustrés sur la figure 5.7 présentant les distances moyennes perçues ainsi que les intervalles de confiance à 95 % associés.

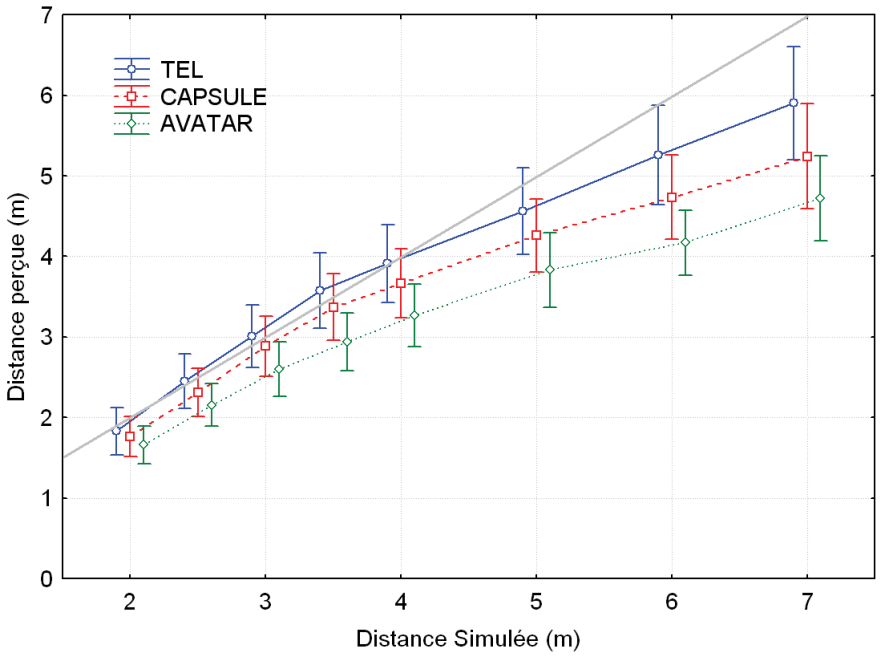


FIGURE 5.7 – *Distances perçues et intervalles de confiance à 95 % des stimuli visuels "Tel" (en bleu), "Capsule" (en rouge) et "Avatar" (en vert).*

La figure 5.7 montre que les participants sont capables d'estimer la distance des différentes conditions ($F(7,161)=193.15$, $p<0.001$), mais que les distances des objets visuels sont globalement sous-estimées, ce qui est en accord avec les résultats issus de la littérature [Teghtsoonian and Teghtsoonian, 1969]. Des différences sont constatées en fonction de l'objet à localiser. L'objet "Avatar" est perçu comme étant plus proche que "Capsule", alors que le stimulus "Tel" est perçu comme étant l'objet le plus lointain. Cette hiérarchie est respectée pour chaque distance même si les différences ne sont pas toujours statistiquement significatives (notamment entre les stimuli "Capsule" et "Tel"). Cet effet du stimulus est confirmé par l'analyse ANOVA ($F(2,46)=16.88$, $p<0.001$) et pourrait être expliqué par la taille relative des objets présentés [Cutting and Vishton, 1995]. Comme pour l'expérience réalisée sur la modalité auditive, les intervalles de confiance montrent une augmentation de la variabilité des jugements en fonction de la distance. Le tableau 5.4 présente les valeurs des coefficients k et a du modèle de Stevens utilisé précédemment.

Stimulus	*k*	*a*	R^2
Tel	1.07	0.90	0.981
Capsule	1.08	0.84	0.972
Avatar	1.03	0.80	0.982

TABLE 5.4 – *Coefficients des fonctions de compression et coefficients de détermination des trois stimuli sonores dans le cas d'objets visuels 3D.*

Au vu des valeurs élevées du coefficient de détermination ($R^2 > 0,97$), la loi de Stevens semble être un bon modèle pour décrire la distance perçue des objets visuels présentés grâce au vidéoprojecteur 3D. Les résultats mettent également en évidence les différences de compression de la distance entre les stimuli. La compression est par exemple plus importante pour le stimulus "Avatar" ($a = 0,80$). Le facteur k est quant à lui relativement stable, indépendamment du stimulus, avec des valeurs légèrement supérieures à 1.

5.4 Perception de la distance d'un objet audio-visuel

Dans cette troisième expérience, les participants doivent estimer la distance d'objets virtuels présentés en condition bimodale, c'est-à-dire des stimuli audio-visuels. Nous choisissons ici de synthétiser les objets sonores et les objets visuels à la même distance. On parle alors de stimuli audio-visuels co-localisés. L'idée est de savoir si l'apport des informations sonores influence la perception de la distance des objets visuels et *vice versa*. En effet, il est possible que la présentation d'informations bimodales congruentes modifie la perception de la distance des objets virtuels en rendant, par exemple, la tâche de localisation plus simple. Ce phénomène se traduirait alors par des intervalles de confiance réduits en comparaison des résultats obtenus précédemment pour les présentations unimodales.

La présentation d'objets audio-visuels impose la prise en compte des contraintes liées à la fois à la présentation visuelle et sonore. Aussi, afin de garantir une visualisation confortable, les conditions de distance testées dans cette expérience sont les mêmes que pour l'expérience sur la présentation visuelle (voir tableau 5.2). Les objets audio-visuels sont donc simulés entre 2 et 7 m. Le fonctionnement du vidéoprojecteur entraîne inévitablement une augmentation du niveau du bruit de fond dans la salle d'expérimentations. Un traitement acoustique a alors été mis en place autour de cet équipement, ce qui limite le niveau de bruit ambiant à 35 $dB(A)$.

Les trois stimuli audio-visuels "Avatar/Voix", "Capsule/Bruit" et "Tel/Sonnerie" sont présentés à un panel de 24 participants (14 femmes et 10 hommes), dont la moyenne d'âge est d'environ 32 ans.

Résultats

La figure 5.8 représente les distances moyennes perçues et les intervalles de confiance à 95 % associés pour les objets audio-visuels simulés.

Ici encore, bien que les participants arrivent à distinguer les différentes conditions de distance ($F(7,161)=118.42$, $p<0.001$), les distances des objets audio-visuels sont sous-estimées. La figure 5.8 illustre également l'augmentation de la variabilité des réponses avec la distance des objets. Cependant, la variabilité des jugements est du même ordre de grandeur que pour la modalité visuelle seule, avec des intervalles de confiance équivalents à environ 20 % de la distance de l'objet simulé. Ce point sera plus largement discuté dans la section 5.5.

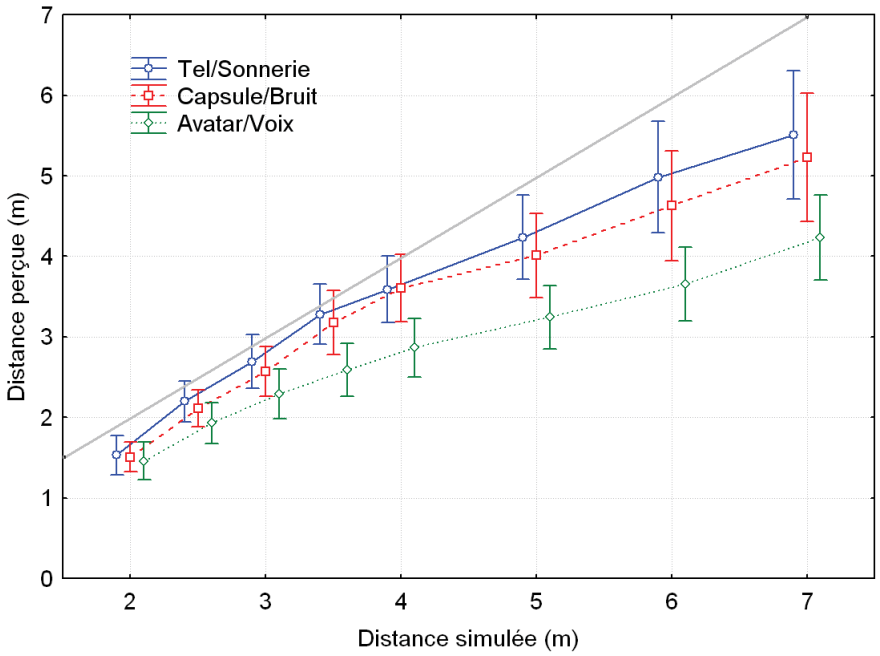


FIGURE 5.8 – *Distances perçues et intervalles de confiance à 95 % des objets audio-visuels co-localisés "Tel/Sonnerie" (en bleu), "Capsule/Bruit" (en rouge) et "Avatar/Voix" (en vert).*

L'analyse ANOVA révèle un effet du stimulus ($F(2,46)=37.34$, $p<0.001$) principalement dû au stimulus "Avatar/Voix" qui est perçu comme étant plus proche que les autres stimuli. En effet, un test *post-hoc* de type Bonferroni révèle qu'il n'y a pas de différences statistiquement significatives entre les deux autres stimuli. Cette sensation de proximité perçue pour le stimulus "Avatar/Voix" semblait prévisible puisqu'un effet comparable a été constaté pour les présentations unimodales. Il est probable que l'effet du stimulus constaté en condition bimodale résulte simplement des effets conjugués observés précédemment.

Le tableau 5.5 présente les valeurs des coefficients k et a du modèle de Stevens dans le cas de cette nouvelle expérience.

Stimulus	*k*	*a*	R^2
Tel/Sonnerie	0.87	0.98	0.974
Capsule/Bruit	0.87	0.95	0.968
Avatar/Voix	0.91	0.80	0.982

TABLE 5.5 – *Coefficients des fonctions de compression et coefficients de détermination des trois stimuli sonores dans le cas d'objets audio-visuels présentés sur le système AV3D.*

Comme précédemment, les valeurs élevées du coefficient de détermination indiquent que la loi de Stevens semble être une bonne approximation de la distance perçue des objets audio-visuels. La compression de la distance est plus importante pour le stimulus "Avatar/Voix"

($a = 0,80$), et le facteur k est stable quel que soit le stimulus audio-visuel présenté.

5.5 Discussions

Nous proposons dans cette section de mettre en relation les résultats issus des expériences réalisées dans ce chapitre sur la perception de la distance. Les deux expériences portant sur la présentation unimodale seront notées A (pour la modalité auditive) et V (pour la modalité visuelle), alors que l'expérience sur la présentation bimodale d'objets co-localisés sera notée AV. Il semble que les effets d'interaction entre modalité et distance ne dépendent pas du stimulus évalué. C'est pourquoi la figure 5.9 représente les résultats des expériences A, V, et AV, moyennés sur tous les stimuli.

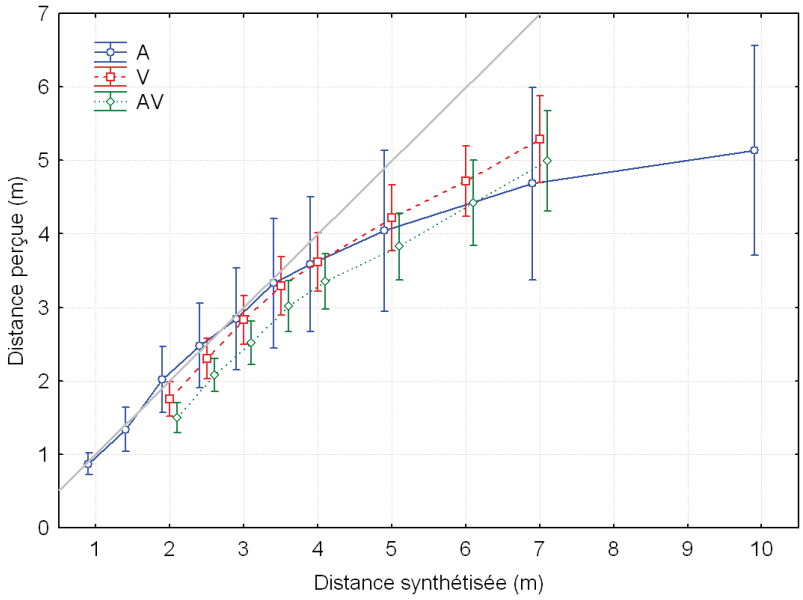


FIGURE 5.9 – *Distances perçues et intervalles de confiance à 95 % des expériences A (en bleu), V (en rouge) et AV (en vert), moyennés sur tous les stimuli.*

Il est impossible de réaliser une analyse ANOVA pour comparer les résultats issus de ces trois expériences, car les conditions de distances diffèrent suivant les expériences (voir tableau 5.2). Cependant, il est possible de commenter les tendances générales qui s'en dégagent (profils de compression, variabilité des jugements, etc.). Les expériences V et AV utilisent quant à elles les mêmes conditions de distance et seront donc comparées au moyen d'une analyse ANOVA. Enfin, les performances de prédiction de la loi de Stevens seront comparées à celles d'un modèle logarithmique.

Présentation sonore contre présentation visuelle

Il est intéressant de comparer les résultats issus des présentations unimodales (modalités auditive et visuelle). Des différences notables ressortent de la comparaison des courbes notées A et V de la figure 5.9.

Pour commencer, la variabilité semble plus importante pour la modalité audio. En effet, les intervalles de confiance représentent environ 30 à 60 % de la distance de la source sonore simulée (valeur moyenne de $43,7\ \% \pm 10,1$), contre environ 20 à 25 % pour la présentation visuelle (valeur moyenne de $21,7\ \% \pm 3,8$).

Ensuite, le phénomène de compression de la perception de la distance semble plus marqué pour la modalité audio seule que pour la modalité visuelle seule. La comparaison des valeurs de l'exposant a dans les tableaux 5.3 et 5.4 va dans ce sens : a est compris entre 0,75 et 0,82 pour les cibles sonores, et entre 0,80 et 0,90 pour les cibles visuelles.

Ces résultats semblent confirmer que la perception de la distance d'un objet est plus performante, à la fois en termes d'erreur de localisation et de variabilité des jugements, avec la modalité visuelle qu'avec la modalité auditive seule.

Présentation visuelle contre présentation audio-visuelle

La figure 5.9 met en évidence les similitudes entre les résultats des expériences V et AV, notamment en termes de compression de la distance perçue. Il apparaît, cependant, que la distance perçue des objets audio-visuels co-localisés a tendance à être sous-estimée par rapport aux jugements basés sur la perception visuelle seule. Une analyse ANOVA est réalisée sur les résultats de ces deux expériences, en considérant les facteurs stimuli et distance simulée. Cette analyse révèle qu'il n'y a pas d'effet statistiquement significatif de la modalité entre ces deux expériences ($F(1,\ 23)=1.84$, $p=0.19$). Les intervalles de confiance des jugements de distance pour les objets audio-visuels correspondent à environ 20 % de la distance de l'objet à localiser (valeur moyenne de $20,5\ \% \pm 2,6$). Ces valeurs sont très légèrement inférieures à celles reportées pour les objets visuels (20 à 25 % de la distance de la cible). Les intervalles de confiance sont détaillés sur la figure 5.10 en fonction de la distance et des stimuli dans les deux expériences.

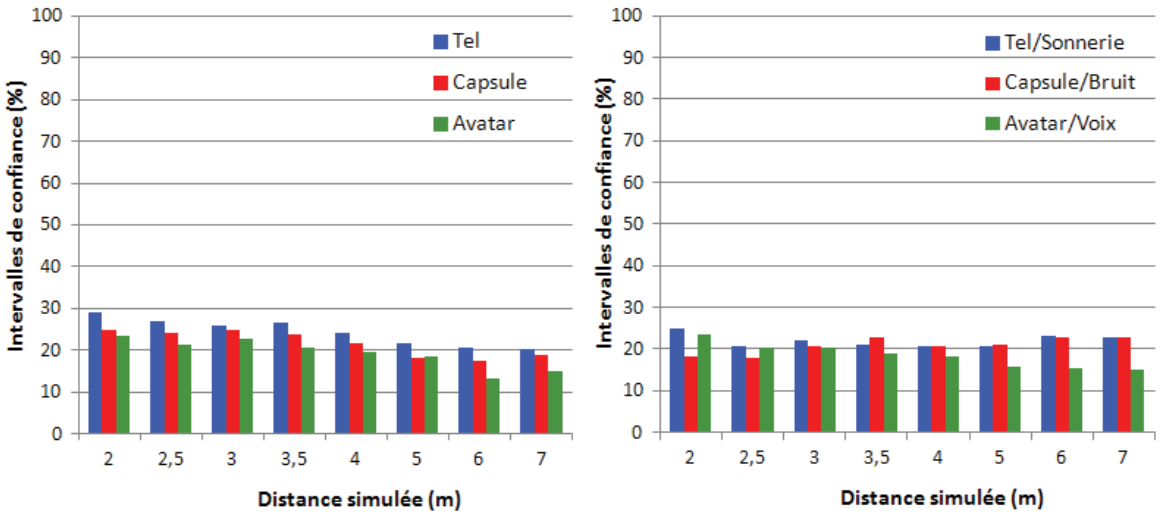


FIGURE 5.10 – *Intervalles de confiance à 95 % exprimés en pourcentage de la distance de l'objet simulé pour la modalité visuelle seule (à gauche) et la présentation audio-visuelle (à droite).*

Effet d'apprentissage

L'étude des intervalles de confiance dans les différentes expériences présentées dans ce chapitre a montré une nette diminution de la variabilité des jugements entre la première expérience sur la modalité auditive (30 à 60 % de la distance simulée) et les expériences portant sur une présentation visuelle (20 à 25 % de la distance simulée) ou bimodale ($\simeq$ 20 % de la distance simulée).

Il faut noter que les différentes expériences ont été réalisées à des dates différentes et que certains participants ont pris part à plusieurs de ces expériences. Aussi, il est naturel de se questionner sur l'attribution de cette augmentation des performances de localisation à un effet d'apprentissage. Les expériences portant sur la modalité visuelle et sur la présentation d'objets audio-visuels co-localisés ont été conduites à un mois d'intervalle et partagent 20 participants sur le panel total de 24. Il y a donc fort à penser qu'un effet d'apprentissage pourrait causer une diminution de la variabilité des résultats entre ces deux expériences. Aussi, aucune conclusion n'est tirée quant à la précision de la perception bimodale par rapport à une présentation unimodale (visuelle, par exemple).

Cependant, il est peu probable qu'un effet d'apprentissage se soit produit entre l'expérience sur la modalité auditive et celle sur la modalité visuelle. En effet, seuls 5 sujets sur 24 ont participé à ces deux expériences qui se sont déroulées à quatre mois d'intervalle. La conclusion avancée précédemment suivant laquelle la perception de la distance d'un objet est plus performante, en termes de variabilité des jugements, avec la modalité visuelle que dans le cas de la modalité auditive seule, n'est donc pas remise en cause.

Modèles de prédiction et loi de Stevens

La distance perçue d'un objet sonore ou visuel est, en général, décrite par la loi de puissance de Stevens (voir section 5.2.2). Les résultats expérimentaux présentés dans cette section ont été comparés à deux lois de comportement (modèles de prédiction) : la loi de Stevens définie par la relation $d_p = k.{d_r}^a$, et un modèle logarithmique du type $d_p = A.ln(d_r) + B$. Dans ces équations, les modèles de prédiction sont caractérisés par deux paramètres : a et k pour la loi de Stevens, et A et B pour le modèle logarithmique. Le tableau 5.6 répertorie les coefficients de détermination R^2 associés à chaque stimulus dans les expériences A (modalité auditive), V (modalité visuelle), et AV (présentation audio-visuelle).

D'après les valeurs du tableau 5.6, les coefficients de détermination du modèle logarithmique sont toujours supérieurs aux valeurs calculées pour la loi de Stevens. Cette observation semble indiquer que, dans le cas de notre étude, le modèle logarithmique est plus adapté pour prédire la perception de la distance d'un objet virtuel (sonore, visuel, ou audio-visuel) que la loi de Stevens. La figure 5.11 propose à titre illustratif les valeurs expérimentales pour le stimulus "Capsule" (expérience sur la modalité visuelle seule), ainsi que les courbes de tendances des deux modèles comparés. Il apparaît clairement sur cette figure que le modèle logarithmique propose une approximation plus précise des valeurs expérimentales que la loi de Stevens.

Expé.	*Stimulus*	*R^2 loi de Stevens*	*R^2 modèle logarithmique*
A	Sonnerie	0.919	0.983
	Bruit	0.935	0.993
	Voix	0.945	0.987
V	Tel	0.981	0.996
	Capsule	0.972	0.999
	Avatar	0.982	0.998
AV	Tel/ Sonnerie	0.974	0.997
	Capsule/ Bruit	0.968	0.995
	Avatar/ Voix	0.982	0.991

TABLE 5.6 – *Coefficients de détermination des modèles de prédiction basés sur la loi de Stevens et sur un modèle logarithmique pour chaque stimulus testé dans les expériences A, V et AV.*

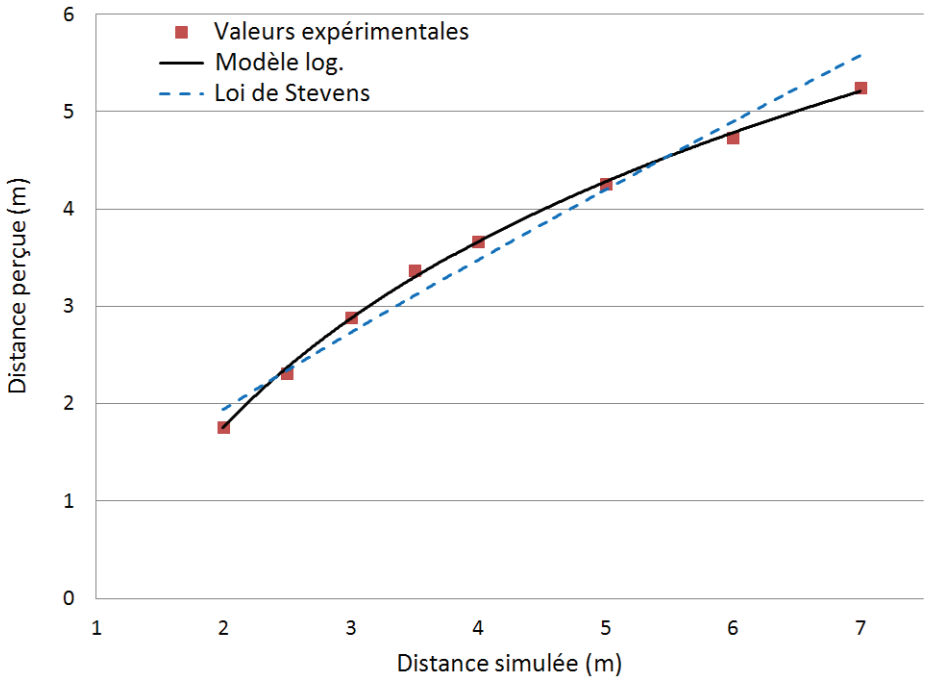


FIGURE 5.11 – *Distance perçue pour le stimulus "Capsule" dans l'expérience portant sur la modalité visuelle et les courbes de tendances associées (R^2 loi de Stevens =0.972, et R^2 modèle logarithmique=0.999, voir tableau 5.6).*

5.6 Conclusion

L'objectif de ce chapitre était de déterminer si le système de restitution audio-visuelle 3D proposé dans le chapitre 4 est capable d'offrir aux spectateurs la sensation de percevoir des objets virtuels sonores et/ou visuels comme étant placés à différentes distances. Pour cela, différentes expériences subjectives ont été menées. Il existe de nombreuses méthodes expérimentales permettant de recueillir des jugements de distance. Dans le cadre de ce travail, nous avons utilisé une méthode relativement simple à mettre en œuvre : la méthode de report direct. Un protocole expérimental a été proposé sur la base de cette méthode de report et a été utilisé dans trois expériences portant sur la perception de la distance d'objets virtuels présentés en conditions unimodale et bimodale.

Les résultats de ces expériences montrent que le système AV3D adopté, associant la WFS à un rendu stéréoscopique, est capable de restituer des informations audio et/ou visuelles suivant la distance. Les participants ont en effet perçu les objets virtuels simulés devant et derrière le dispositif de diffusion utilisé. La perception de la distance des stimuli sonores, visuels et audio-visuels est généralement sous-estimée avec un phénomène de compression plus ou moins marqué suivant la modalité impliquée. La perception de la distance des objets sonores, par exemple, semble plus compressée et également plus variable que dans le cas d'objets visuels ou audio-visuels. Ensuite, les résultats issus des expériences portant sur les stimuli visuels et audio-visuels montrent des comportements globalement similaires en termes de variabilité et de niveau de compression de la perception. Un effet du stimulus a été mis en évidence dans chacune des expériences : le stimulus "Avatar" et/ou "Voix" est perçu comme étant le plus proche, quelle que soit la modalité étudiée.

Il a donc été montré que le système AV3D est capable de simuler des objets sonores et/ou visuels perçus comme étant placés à différentes distances, devant et derrière le dispositif de restitution. Les expériences portant sur la présentation unimodale ont montré qu'un objet virtuel peut être perçu à des positions différentes suivant la modalité stimulée (audition ou vision). Il est possible que ces écarts de localisation aient un impact sur la perception de la distance d'objets présentés en condition bimodale. Or, la nature et l'ampleur des phénomènes d'interaction multimodale suivant la distance (effet ventriloque et ventriloque inverse, intégration spatiale, etc.) ont relativement peu été explorées jusqu'à présent. Pour cette raison, nous proposons d'étudier les phénomènes d'interaction audio-visuelle dans le chapitre 6, à travers une expérience portant sur la perception de la distance de stimuli audio-visuels spatialement incohérents. Il sera également question de déterminer les écarts pour lesquels des stimuli audio et visuels disparates sont intégrés par notre système perceptif. Ce phénomène d'intégration aboutit à la compréhension d'un percept audio-visuel unique par les spectateurs.

Chapitre 6

Perception d'objets audio-visuels spatialement incohérents et intégration multimodale

6.1 Introduction

Les expériences présentées dans le chapitre 5 ont porté sur la perception de la distance d'objets sonores, visuels et audio-visuels. Dans ce dernier cas, les informations sonores et visuelles ont été simulées à la même distance (objets co-localisés). Or, nous avons pu observer qu'un objet virtuel peut être perçu à des distances différentes suivant la modalité stimulée (audition ou vision). Ainsi, il est possible que ces écarts aient engendré la perception d'incohérences, en termes de localisation en distance, lors de la présentation bimodale. D'une manière plus générale, cette réflexion soulève la problématique de l'influence des incohérences spatiales entre les stimuli sonores et visuels sur la perception audio-visuelle (bimodale).

Nous proposons dans ce chapitre d'étudier les phénomènes d'interaction multimodale qui peuvent être engendrés par la présentation d'objets audio-visuels spatialement incohérents suivant la distance. En effet, des expériences décrites dans la littérature (section 1.4.1) montrent que la localisation d'un stimulus sonore peut, sous certaines conditions, être biaisée par la présence d'un stimulus visuel et *vice versa*. D'autre part, notre système perceptif semble capable d'intégrer des percepts unimodaux en tolérant certains écarts entre la position des stimuli sonores et visuels (voir section 1.4.3).

Ces exemples d'interactions audio-visuelles sont étudiés au moyen de deux expériences subjectives. Dans la première, il est question d'explorer l'influence des incohérences audio-visuelles sur la perception de la distance. La deuxième expérience est, quant à elle, réalisée dans le but d'estimer les zones d'intégration audio-visuelle en distance. Cette dernière étude nous permettra notamment de déterminer à quelles distances les sources sonores peuvent être simulées pour être perçues comme spatialement cohérentes avec un objet visuel cible.

6.2 Perception de la distance d'un objet audio-visuel spatialement incohérent

6.2.1 Motivations

Nous souhaitons, dans cette expérience, présenter aux participants des objets virtuels audio-visuels non plus co-localisés, mais présentant des disparités en termes de distance. Le but de cette expérience est de savoir si la présentation d'informations sonores spatialement incohérentes peut biaiser la localisation d'un objet visuel suivant la distance. Ce phénomène, appelé effet ventriloque-inverse [Alais and Burr, 2004] (voir section 1.4.1), impliquerait qu'un objet visuel serait perçu comme étant plus proche, lorsqu'un objet sonore est placé entre l'observateur et l'objet visuel. Au contraire, si l'objet sonore est placé derrière l'objet visuel, alors la distance perçue de ce dernier augmenterait. Si cet effet d'attraction du stimulus visuel par le stimulus sonore est observé, il serait alors possible de considérer la restitution de la distance sonore comme une solution potentielle pour améliorer le confort de visualisation de contenus 3D, ou l'expérience audio-visuelle 3D. En effet, ce résultat suggérerait qu'il est possible d'augmenter la sensation de relief visuel perçu sans pour autant augmenter les disparités à l'écran, mais uniquement grâce à la restitution sonore. Avec le même raisonnement, il serait possible de maintenir le niveau de relief visuel perçu grâce au son, tout en diminuant les disparités visuelles qui sont potentiellement responsables de phénomènes d'inconfort.

6.2.2 Protocole expérimental

Dans cette expérience, la tâche des sujets consiste à évaluer la distance des objets visuels, placés entre 2 et 7 mètres comme dans les expériences précédentes (chapitre 5), en présence de sons "distracteurs" placés à 2, 3 ou 5 mètres. Ainsi, chacune des huit conditions de distance visuelle est présentée avec trois conditions sonores. Il en résulte un nombre total de 24 conditions audio-visuelles par stimulus.

Comme dans les expériences précédentes, les trois stimuli audio-visuels "Avatar/Voix", "Capsule/Bruit" et "Tel/Sonnerie" sont présentés aux participants dans des sessions différentes, dont l'ordre de présentation est alterné.

Nous choisissons de mettre en place le même protocole expérimental que celui utilisé précédemment (décrit dans la section 5.2.3), car cette expérience concerne également la localisation en distance d'objets virtuels. Ainsi, les jugements sont recueillis par report direct au clavier. Notons cependant que le nombre de répétitions des conditions audio-visuelles est ramené à trois au lieu de cinq. Ceci permet de limiter le temps de passation du test sans pour autant altérer la variabilité des résultats [1].

L'expérience est réalisée dans les mêmes conditions expérimentales que précédemment (distance de visualisation, taille de l'image, niveau sonore, bruit de fond, etc.). Le test est réalisé par un panel constitué de 24 participants (14 femmes et 10 hommes), dont la moyenne d'âge est d'environ 31 ans.

1. Une analyse a été menée sur les résultats des expériences présentées dans le chapitre 5. Les résultats calculés sur les trois premières répétitions sont comparables, en termes de moyennes et d'intervalles de confiance, aux résultats d'origine impliquant cinq répétitions.

6.2.3 Résultats

Nous avons choisi de réaliser une analyse ANOVA pour chaque stimulus audio-visuel, en considérant deux facteurs : la distance de l'objet visuel et la distance de l'objet sonore qui l'accompagne. Ici, l'influence du stimulus sur les jugements de distance n'est pas étudiée, car ce point a déjà été discuté dans les expériences précédentes.

Les analyses ANOVA révèlent que les jugements de distance ne sont pas influencés par la distance de l'objet sonore pour les stimuli "Avatar/Voix" ($F(2,46)=0.34$, $p=0.71$) et "Capsule/Bruit" ($F(2,46)=0.46$, $p=0.57$). En revanche, pour le stimulus "Tel/Sonnerie", un léger effet de la distance des objets sonores est observé ($F(2,46)=3.78$, $p<0.03$). La figure 6.1 reproduit les résultats de cette expérience pour le stimulus audio-visuel "Tel/Sonnerie".

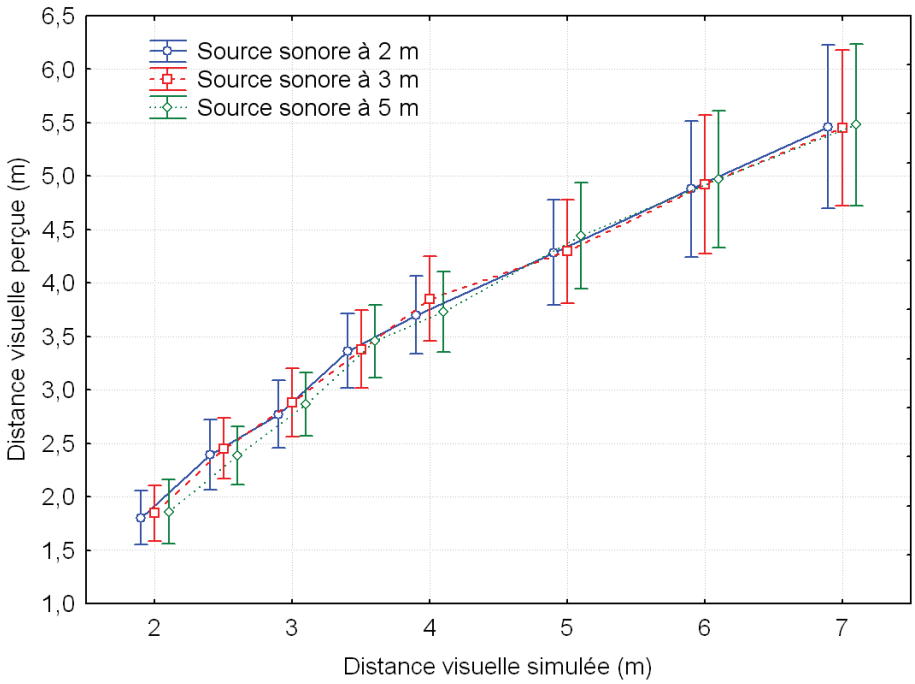


FIGURE 6.1 – *Distances visuelles perçues et intervalles de confiance à 95 % pour l'objet "Tel/Sonnerie" lorsque la source sonore est placée à 2 m (en bleu), 3 m (en rouge) ou 5 m (en vert).*

La figure 6.1 montre que les distances moyennes semblent comparables entre les conditions sonores testées. Il est donc difficile de commenter l'effet de la condition audio révélée par l'analyse ANOVA dans le cas du stimulus "Tel/Sonnerie" [2]. Compte tenu de ces résultats, il n'est pas possible de conclure sur la présence d'un éventuel effet ventriloque inverse. Il est à noter que la variabilité des jugements est de l'ordre de 20 % de la distance de l'objet visuel simulé comme dans le cas des objets audio-visuels co-localisés (voir section 5.4).

2. Il a été choisi de ne pas présenter les résultats pour les stimuli "Avatar/Voix" et "Capsule/Bruit" dans ce document, car il n'y a pas de différences observables entre les trois conditions sonores testées.

6.2.4 Discussions

Les résultats de cette expérience n'ont donc pas permis de mettre en évidence un effet ventriloque inverse. Il est possible que l'utilisation d'un protocole expérimental différent, avec une méthode de recueil adaptée pour limiter la variabilité des estimations, aurait pu mettre en évidence un effet de capture par la modalité auditive. De plus, dans le but de tester l'influence de la distance sonore sur la perception de la distance visuelle, les participants ont eu pour consigne d'estimer la distance de l'objet visuel. Il est probable que des consignes indiquant d'estimer la distance des objets audio-visuels auraient amené des résultats différents, avec une plus forte influence de la condition sonore. Toutefois, il est également possible que la saillance des informations visuelles soit trop élevée pour que ces dernières soient biaisées par les informations sonores. Ces hypothèses n'ont pas fait l'objet de recherches plus approfondies dans le cadre de ce travail.

Cette expérience a montré que la présence d'objets sonores placés à des positions arbitraires, et potentiellement incohérentes par rapport aux objets visuels, ne semble pas avoir d'incidence sur la perception de la distance des objets visuels. Ce résultat nous amène à nous interroger quant à la sensibilité des participants à détecter les incohérences spatiales audio-visuelles. En effet, il est fort probable que de telles incohérences soient tolérées par notre système perceptif lorsque les écarts de localisation entre les stimuli sonores et visuels sont inférieurs à un certain seuil. Des phénomènes similaires ont été mis en évidence auparavant dans différentes études portant sur des tâches de localisation en azimut (voir section 1.4.3). En revanche, nous pouvons nous attendre à la perception d'incohérences dès lors que les disparités en distance sont plus élevées. Cependant, les données recueillies dans cette expérience ne permettent pas de remonter à ces informations, relatives à la notion de cohérence spatiale. Pour cette raison, nous proposons, à travers l'expérience décrite dans la section suivante, d'explorer les écarts de localisation à ne pas dépasser afin de maintenir la perception d'éléments audio-visuels cohérents.

6.3 Estimation des zones d'intégration audio-visuelle en distance

Il a été montré que le système de restitution audio-visuelle AV3D permet de simuler des objets virtuels sonores et/ou visuels perçus à différentes distances. Cependant, rien ne garantit que les espaces ainsi restitués (distance sonore et distance visuelle) soient perçus par les utilisateurs comme étant cohérents. Nous savons que, sous certaines conditions, la plasticité de notre cerveau permet d'intégrer des informations multimodales présentant des disparités en termes de localisation. Ces conditions d'intégration peuvent porter sur les écarts de localisation entre les stimuli unimodaux ou sur la nature même des stimuli (voir section 1.4.3). Dans le contexte de cette étude, nous nous intéressons plus particulièrement aux disparités spatiales suivant la distance.

L'objectif de cette expérience est de déterminer les écarts de localisation en distance entre un objet sonore et un objet visuel, pour lesquels les spectateurs perçoivent un objet audio-visuel cohérent. Ces écarts couvrent une zone de l'espace appelée zone de fusion (ou zone d'intégration audio-visuelle). Dans cette expérience, les stimuli audio-visuels sont présentés grâce au système de restitution AV3D.

6.3.1 Protocole expérimental pour l'estimation des zones d'intégration

Dans cette expérience, les objets sonores et visuels sont simulés au même endroit en termes d'azimut et d'élévation, mais à des distances différentes. Les cibles visuelles sont placées à quatre distances : 2, 3, 4, et 5 m. Pour chaque position de l'objet visuel, dix sources sonores synthétisées par WFS sont simulées entre 1 et 10 m (1, 1.5, 2, 2.5, 3, 3.5, 4, 5, 7, et 10 m). Au total, les 40 conditions audio-visuelles présentées offrent des informations bimodales spatialement cohérentes dans 10 % des cas. La figure 6.2 illustre la position des objets virtuels simulés devant et derrière le système de restitution AV3D.

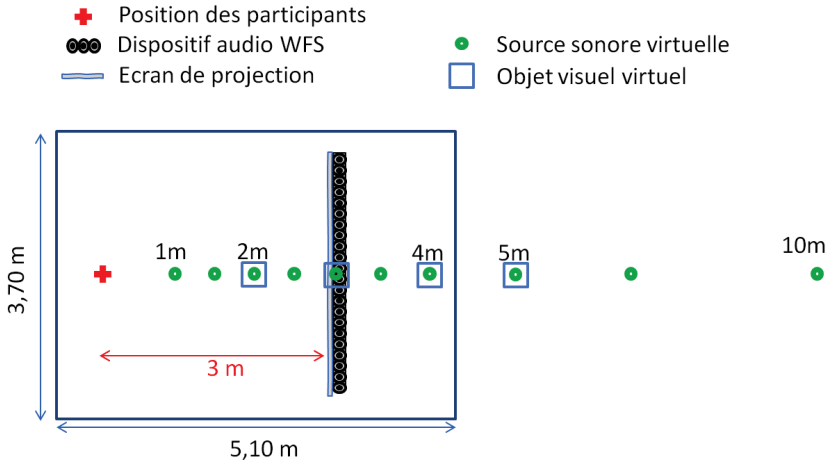


FIGURE 6.2 – *Position des objets visuels (entre 2 et 5 m) et des objets sonores (entre 1 et 10 m) présentés.*

Après chaque présentation, les participants estiment la cohérence spatiale des objets audio-visuels en indiquant, par un vote de type "oui/non", si le son est placé au même endroit que l'objet visuel.

Les trois stimuli audio-visuels décrits dans la section 5.2.3 sont testés afin d'étudier l'influence potentielle du stimulus sur l'intégration audio-visuelle suivant la distance. Le test est divisé en trois sessions (une session par type de stimulus). L'ordre de présentation des sessions est alterné de manière à présenter toutes les combinaisons possibles (sessions 1/2/3, ou 1/3/2, ou 2/1/3, etc.) au même nombre de participants. Lors de chaque session, les 40 conditions audio-visuelles sont présentées 5 fois, le tout dans un ordre aléatoire. Les jugements de cohérence sont recueillis après la présentation de chaque essai. Les stimuli sont présentés en continu jusqu'au jugement des participants.

Une phase de familiarisation permet aux participants de découvrir un échantillon des conditions à évaluer [3] et de prendre en main l'interface de test.

L'expérience est conduite sur un panel constitué de 24 participants (14 femmes et 10 hommes) dont la moyenne d'âge est de 32 ans. Les sujets ont, pour la plupart, l'habitude de réaliser des tests d'écoute et de visionner des contenus vidéo 3D. Cependant, ces personnes n'ont jamais été amenées à juger la cohérence spatiale de stimuli audio-visuels.

6.3.2 Résultats

La proportion d'intégration audio-visuelle est calculée pour chaque participant sur la base des 5 réponses "oui" ou "non" attribuées à chaque condition audio-visuelle. Les jugements de type "oui/non" sont préalablement convertis en valeurs numériques (1 pour "oui" ou 0 pour "non"). La proportion d'intégration audio-visuelle est ainsi définie comme la proportion de réponses "oui" (100 % de réponses "oui" correspondant à 1).

Une analyse de la variance (ANOVA) est réalisée à partir des proportions individuelles d'intégration audio-visuelle, en considérant deux facteurs : le stimulus audio-visuel testé et la distance de l'objet sonore. Les résultats de cette analyse ANOVA montrent que, pour les quatre distances visuelles, la proportion d'intégration dépend de la position des objets sonores simulés de manière significative ($p<0.001$). En d'autres termes, ces résultats démontrent que les participants sont sensibles à la cohérence spatiale des conditions audio-visuelles à l'essai.

Les résultats de cette expérience sont présentés sur la figure 6.3. Ces graphiques représentent les proportions de conditions jugées spatialement cohérentes en fonction de la distance de l'objet sonore, pour les quatre conditions visuelles testées (2, 3, 4, et 5 m). Les proportions d'intégration audio-visuelles reportées sont exprimées entre 0 et 1 : 0 si les stimuli audio et visuels sont perçus comme étant spatialement incohérents (aucun jugement "oui"), et 1 s'ils sont jugés spatialement cohérents (100 % de jugements "oui").

3. L'échantillon sélectionné est jugé comme étant représentatif de l'ensemble des conditions testées.

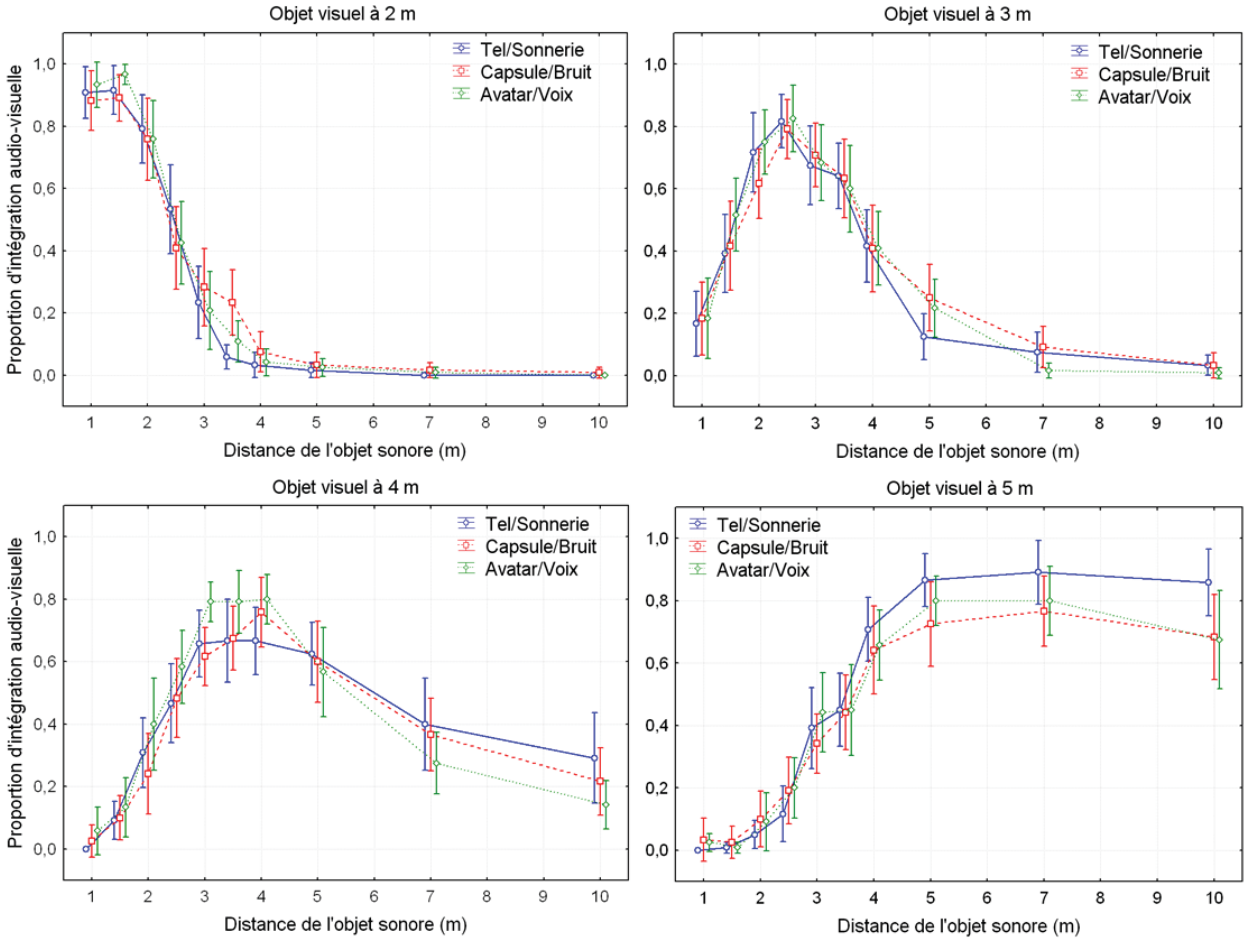


FIGURE 6.3 – *Proportions de conditions jugées spatialement cohérentes en fonction de la distance simulée de l'objet sonore. Les quatre graphiques correspondent aux quatre distances simulées des objets visuels : 2, 3, 4 et* 5 *m.*

Nous pouvons observer sur la figure 6.3 que les notes de cohérence sont réparties entre 0 et plus de 0,8 pour chacune des distances visuelles testées, avec un maximum de 0,96 pour l'objet visuel "Avatar" placé à 2 m et le son "Voix" simulé à $1,5$ m. Cette dynamique importante entre les notes moyennes minimales et maximales souligne la pertinence du choix des conditions à l'essai. En effet, pour chaque distance visuelle évaluée, certaines conditions sonores sont considérées comme étant spatialement cohérentes, et d'autres comme étant incohérentes. Ces données permettront de déterminer ultérieurement les seuils de tolérance aux incohérences spatiales.

Il est important de rappeler que la distance du système de restitution audio-visuelle est fixée à 3 m. Par conséquent, l'objet sonore synthétisé à 3 m correspond à une source virtuelle placée juste au niveau de l'écran. Cette position de la source sonore par rapport à l'écran de projection est équivalente à la configuration classiquement rencontrée dans les installations

destinées au grand public [4]. Si on considère les estimations de cohérence spatiale propres à cette position de source sonore (3 m), des incohérences spatiales semblent être perçues notamment pour l'objet visuel placé à 2 m. En effet, les proportions d'intégration pour les trois stimuli sont faibles, avec des valeurs comprises entre 0,2 et 0,3 (soit 20 à 30 % des réponses). Cette observation semble conforter notre hypothèse de travail suivant laquelle les systèmes de restitution audio-visuelle 3D actuellement utilisés pour le grand public sont susceptibles d'entraîner la perception d'incohérences, en termes de localisation en distance, entre les stimuli audio et visuels.

Même si de légères variations sont observables localement entre les trois stimuli (objets visuels à 4 ou 5 m notamment), il est cependant difficile de les différencier. L'ANOVA confirme l'absence d'effet significatif quant à l'influence potentielle du type de stimulus sur la cohérence spatiale ($p>0.5$ pour les objets visuels placés à 2 et 3 m, et $p>0.1$ pour les autres distances d'objets visuels). Considérant cette absence d'effet statistique, nous proposons d'étudier les notes moyennes obtenues sur l'ensemble des trois stimuli.

Pour ce faire, les jugements de cohérence spatiale ont été analysés d'après une méthode inspirée de travaux portant sur l'estimation de la simultanéité subjective [Yarrow et al., 2011]. Conformément à la méthode décrite par Yarrow et al., les données expérimentales sont modélisées au moyen de deux gaussiennes. Cette technique permet de prendre en compte, de manière indépendante, les comportements observés avant et après la position de l'objet sonore pour laquelle la proportion d'intégration audio-visuelle est maximale. La dissociation des parties "ascendantes" et "descendantes" de ces courbes rend le modèle flexible, ce qui offre une meilleure approximation des données expérimentales en comparaison d'une modélisation basée sur une courbe gaussienne unique (symétrique). Les graphiques présentés sur la figure 6.3 montrent en effet des profils asymétriques entre les objets sonores perçus devant et derrière la cible visuelle (objet visuel à 4 m, par exemple).

La figure 6.4 présente la modélisation des données expérimentales réalisée grâce à la méthode décrite par Yarrow et al. [Yarrow et al., 2011] pour les quatre distances visuelles.

Comme dans certaines études mentionnées dans la section 1.4.3 ([Gorzel et al., 2012, Corrigan et al., 2013]), nous définissons les limites de la zone d'intégration audio-visuelle en distance par l'estimation des seuils à 50 %. Ainsi, la zone d'intégration est délimitée par les conditions audio-visuelles perçues comme spatialement cohérentes par plus de 50 % des réponses. Les seuils à 50 % sont déterminés grâce aux modélisations illustrées sur la figure 6.4, et sont répertoriés dans le tableau 6.1 avec l'étendue des zones d'intégration.

4. Le système de restitution sonore multicanal 5.1 est probablement le format le plus utilisé dans les salles de cinéma actuelles (voir chapitre 2). Le système 5.1 est notamment constitué de 3 haut-parleurs placés dans la zone frontale avec un canal central placé juste derrière l'écran de projection.

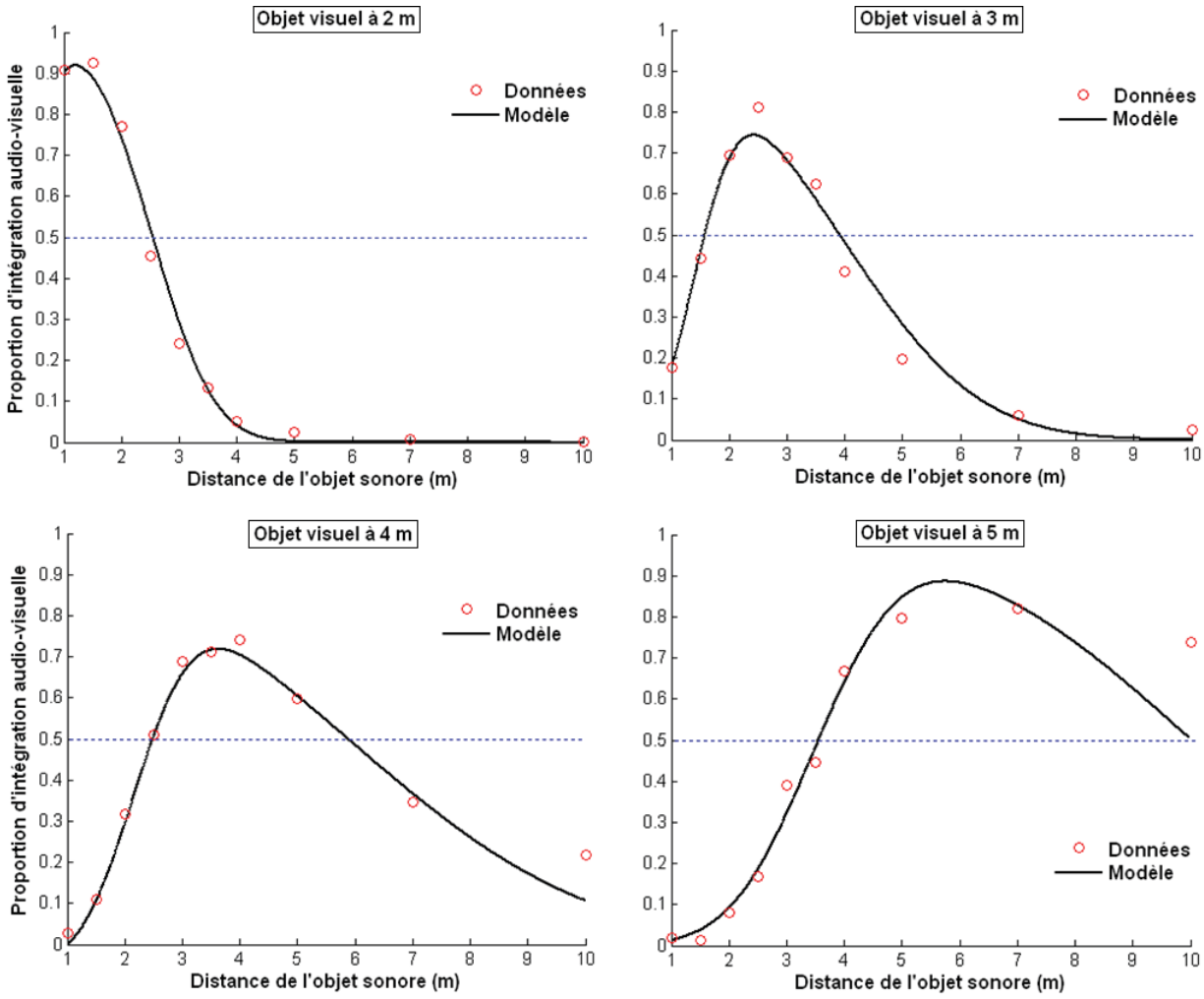


FIGURE 6.4 – *Données expérimentales moyennées sur les trois stimuli et modélisées d'après la méthode décrite dans [Yarrow et al., 2011]. Les quatre graphiques correspondent aux quatre distances des objets visuels : 2, 3, 4 et 5 m.*

Distance de l'objet visuel	*2m*	*3m*	*4m*	*5m*
Seuils à 50 % inférieur et supérieur (en m)	< 1 2.54	1.42 3.91	2.17 5.88	3.49 >10
Etendue de la zone d'intégration (en m)	> 1.54	2.49	3.71	> 6.51

TABLE 6.1 – *Seuils à 50 % et étendue des zones d'intégration estimés à partir de la modélisation des valeurs expérimentales moyennes (voir figure 6.4).*

Notons que les seuils à 50 % ne peuvent pas être déterminés pour les objets visuels placés à 2 et 5 m, car la gamme de distances sonores testée semble insuffisante. Cependant, pour les cibles visuelles placées à 3 et 4 m, les zones d'intégration sont estimées et s'étendent respectivement sur 2,49 et 3,71 mètres. Il est possible de comparer cette dernière valeur à certains travaux issus de la littérature. En effet, Gorzel et al. ont estimé la fenêtre d'intégration pour un objet visuel placé à 4 m des spectateurs, dans le cas d'une reproduction sonore ambisonique [Gorzel et al., 2012]. Leur étude montre que, suivant l'ordre ambisonique considéré, la fenêtre d'intégration audio-visuelle s'étend sur $4,34$ m (ambisonique à l'ordre 1) ou $0,83$ m (ambisonique à l'ordre 0). Notons que la zone d'intégration déterminée dans notre étude (impliquant une reproduction sonore par Wave Field Synthesis) semble relativement cohérente avec les résultats avancés par Gorzel et al. dans le cas de l'ambisonique à l'ordre 1. Cependant, l'étude menée par Gorzel et al. ne permet pas de définir l'étendue des zones d'intégration dans le cas d'objets visuels placés à d'autres distances. Il est donc difficile de commenter davantage une éventuelle similarité entre les résultats obtenus pour ces deux technologies de reproduction sonore.

Le tableau 6.1 permet également d'observer l'évolution de l'étendue des zones d'intégration en fonction de la distance. Il apparaît assez clairement que l'étendue de ces fenêtres a tendance à augmenter avec la distance de l'objet visuel. Il est fort probable que cette extension soit liée au flou de localisation qui augmente lorsque les objets sonores et visuels s'éloignent. Ce point mériterait d'être confirmé en testant une gamme de distances sonores plus importante, ou bien des positions visuelles intermédiaires (2,5, 3,5 et $4,5$ m, par exemple).

6.4 Conclusion

Le but de ce chapitre était d'étudier les phénomènes d'interaction multimodale résultant de la présentation d'objets audio-visuels spatialement incohérents. Pour cela, deux expériences subjectives ont été menées dans lesquelles les positions d'objets sonores et visuels présentent des disparités suivant la distance.

La première expérience portait sur la perception de la distance dans le cas d'objets audio-visuels non co-localisés. Les résultats de cette expérience indiquent que les différentes informations sonores présentées n'ont pas d'incidence sur la distance perçue d'un objet visuel. Dans ce contexte, il n'a pas été possible de mettre en évidence l'existence d'un effet ventriloque inverse. En tout état de cause, les effets recherchés, s'ils existent, sont certainement subtils. Il est donc possible que le protocole expérimental, basé sur une méthode d'estimation verbale, engendre des jugements de distance présentant une variabilité trop élevée pour pouvoir observer ces phénomènes d'interaction audio-visuelle. Une autre explication repose sur la saillance des informations visuelles qui est peut-être trop importante pour laisser place à tout phénomène de capture par la modalité auditive.

Nous avons ensuite proposé d'évaluer la sensibilité des spectateurs aux incohérences spatiales audio-visuelles générées. Nous avons pour cela déterminé les écarts pour lesquels des stimuli audio et visuels placés à différentes distances sont intégrés et fusionnés par le système perceptif. Ce phénomène d'intégration aboutit à la compréhension d'un percept audio-visuel unique lorsque les objets sonores et visuels sont placés dans une zone limitée de l'espace, appelée fenêtre d'intégration. Les résultats de cette expérience montrent que les participants sont capables de détecter des conditions audio-visuelles incohérentes. De plus, les spectateurs semblent davantage tolérants aux écarts de localisation pour les objets visuels les plus éloignés. En effet, l'étendue des fenêtres d'intégration a tendance à augmenter avec la distance des objets visuels. Enfin, cette expérience confirme l'idée que les systèmes de restitution sonore actuellement utilisés pour accompagner les vidéos 3D sont susceptibles d'entraîner la perception d'évènements audio-visuels spatialement incohérents, en particulier pour un objet visuel en jaillissement.

Nous sommes à présent capables de restituer des objets virtuels à différentes distances grâce au système AV3D en connaissant les conditions à respecter, en termes de localisation en distance, pour assurer la perception unifiée des stimuli multimodaux. Nous souhaitons maintenant utiliser l'ensemble de ces données pour savoir si la qualité d'expérience audio-visuelle 3D peut être améliorée en présentant des objets audio-visuels perçus comme étant spatialement cohérents par les spectateurs.

Nous souhaitons également étudier l'impact potentiel de la localisation des stimuli sonores sur la sensation de relief lors de la diffusion d'une scène audio-visuelle. En effet, bien que le phénomène de capture auditive n'ait pas été mis en évidence dans le cas d'une tâche de localisation (section 6.2), il est possible qu'une tâche de plus haut niveau, telle que l'estimation de la sensation de profondeur visuelle, puisse être influencée par la présence d'informations sonores en distance.

Chapitre 7

Impact du rendu de la distance d'objets sonores sur la qualité d'expérience audio-visuelle 3D

7.1 Introduction

Nous avons montré dans le chapitre 5 que les spectateurs sont capables de percevoir la distance d'objets sonores synthétisés par Wave Field Synthesis, et/ou d'objets visuels présentés en 3D. Les spectateurs semblent également sensibles à la cohérence spatiale entre ces informations unimodales (chapitre 6). Ensuite, nous avons déterminé les positions auxquelles les objets virtuels doivent être synthétisés pour garantir la perception d'évènements audio-visuels unifiés. Cette dernière notion semble en effet primordiale pour offrir aux spectateurs une expérience audio-visuelle 3D immersive.

Nous souhaitons à présent savoir si la restitution de la distance d'objets sonores permet d'améliorer la qualité d'expérience des spectateurs dans un contexte d'utilisation réaliste. Pour cela, différentes séquences audio-visuelles naturelles sont captées lors d'un tournage (voir section 7.2). Ces séquences sont ensuite présentées dans une expérience subjective (section 7.3). Plusieurs scènes sonores sont alors testées pour accompagner les images 3D : un mixage sans restitution de la distance sonore, un mixage où la distance sonore est réaliste par rapport à la position visuelle des objets en profondeur et un mixage où la restitution de la distance sonore des objets est volontairement amplifiée. Les résultats de cette expérience sont présentés et commentés à la fin de ce chapitre.

7.2 Captation de séquences audio-visuelles 3D

La première étape de ce travail consiste à présenter aux spectateurs des séquences audio-visuelles réalistes. En effet, dans le but d'évaluer la qualité d'expérience liée à la visualisation de contenus vidéo 3D, il ne suffit pas de présenter un objet de synthèse placé dans un environnement virtuel. La visualisation de scènes complexes constituées d'éléments visuels riches (personnages, décor, etc.) est sans doute plus représentative d'un contexte réel d'utilisation. Pour évaluer l'influence du rendu de la distance sonore sur l'expérience audio-visuelle 3D, nous devons être en mesure de manipuler les composantes sonores de séquences audio-visuelles complexes. Or, le contrôle et le positionnement de chaque source sonore dans l'espace impliquent de disposer de tous les fichiers audio originaux (un fichier par source sonore) [1]. De plus, pour spatialiser ces sources sonores de manière pertinente, il est nécessaire de connaître avec précision la localisation des éléments visuels présents dans la scène.

Pour ces différentes raisons, nous décidons de générer les contenus audio-visuels 3D en réalisant une captation naturelle à l'aide d'une caméra Panasonic AG-3DA1 à double capteurs (voir figure 7.1). La création de contenus permet en effet de respecter les contraintes liées aux besoins de l'expérience (contrôle de la position des éléments visuels et sonores, captation indépendante des sources sonores en vue de leur manipulation). Il faut également prendre en compte les contraintes liées au système de restitution AV3D (disparités maximales, restitution du relief). Cette section détaille ces différentes contraintes et propose une description des séquences audio-visuelles créées.

FIGURE 7.1 – *Caméra Panasonic AG-3DA1 utilisée lors de la captation des séquences 3D.*

1. Une autre méthode consiste à utiliser des fichiers audio multicanaux existants (mixés au format 5.1 par exemple) et d'en extraire les sources sonores à l'aide de méthodes de séparation de sources. Cependant, la qualité de reproduction sonore est variable suivant la méthode de séparation utilisée. Nous souhaitons offrir aux spectateurs des contenus audio de qualité optimale, ce qui explique pourquoi ces méthodes n'ont pas été mises en œuvre dans le cadre de ce travail.

7.2.1 Prise en compte des besoins expérimentaux

L'expérience présentée dans la section 7.3 implique la manipulation de sources sonores dans l'espace, en utilisant comme données d'entrée la position des objets visuels. Il est nécessaire de prendre en compte ces besoins dans la phase de création des contenus.

Captation des sources sonores en vue de leur manipulation

Pour pouvoir manipuler chaque source sonore suivant la distance, il est nécessaire de disposer d'un fichier audio par source. Nous avons donc effectué une prise de son pour chaque source sonore à l'aide de microphones cravates DPA 4080 (microphones à directivité cardioïde) placés sur les acteurs. Des enregistrements monophoniques ont également été réalisés pour les sources sonores ajoutées lors de l'étape de post-production (bruitages). Ces différentes prises de son permettent le transport des informations sonores captées suivant le format de représentation *object-based* également appelé "tout paramétrique" (voir section 2.2.6). Ce format permet la manipulation de chaque source sonore de manière indépendante lors de la phase de diffusion, indépendamment du système de restitution sonore utilisé.

Mesure de la position des éléments visuels

Afin d'offrir aux spectateurs des objets audio-visuels cohérents, le placement des sources sonores doit être réalisé en tenant compte des positions des objets visuels et de l'étendue des zones d'intégration (voir section 6.3). Il est donc important de repérer avec précision, lors de la phase de captation, les positions des objets qui seront visibles à l'écran (distance et azimut).

7.2.2 Prise en compte du système de restitution visuelle cible

Contrôle du relief restitué

Comme mentionné dans la section 2.3.2, la perception du relief dépend à la fois des paramètres de captation et des conditions de visualisation des séquences. Afin de reproduire fidèlement les distances et les dimensions des objets, il est nécessaire d'utiliser un jeu spécifique de paramètres de captation par rapport aux conditions de visualisation. Ainsi, la distance de convergence et l'ouverture des capteurs sont définies de manière à correspondre respectivement à la distance de visualisation (3 m) et à la taille de l'écran dans le champ de vision des spectateurs (image de 2 m de base, soit environ 37° du champ visuel). L'écart inter-caméras est quant à lui équivalent à l'écart interoculaire moyen (65 mm). Ces paramètres assurent que le relief de la scène captée sera restitué de manière linéaire sur le système AV3D, tout comme la dimension des objets (1 m dans la scène captée correspond à 1 m dans la scène restituée). La restitution visuelle respecte alors la condition de conformité totale.

Contrôle des disparités maximales à ne pas dépasser

Il est également important de prendre en considération les disparités maximales à ne pas dépasser lors de la diffusion des contenus vidéo 3D. En effet, si les disparités positives (objets placés derrière l'écran) ou négatives (objets en jaillissement) sont trop importantes, des phénomènes de gêne visuelle et de fatigue peuvent apparaître (voir section 2.3.2). Afin d'éviter

ce problème d'inconfort, les objets visuels de la scène doivent être placés dans une gamme de distance restreinte. Compte tenu du système de restitution utilisé, les objets visuels de la scène doivent être présentés à des distances comprises entre $1,87\ m$ et $7,5\ m$ du spectateur (voir section 4.3.2). Étant donné la condition de conformité totale évoquée précédemment, cette contrainte impose que les objets de la scène captée soient placés dans la même gamme de distance.

7.2.3 Description des séquences audio-visuelles captées

Le tournage des séquences audio-visuelles 3D a été réalisé dans un studio d'Orange Labs avec le support d'une équipe technique spécialisée (cadreur, directeur lumière, ingénieur du son, stéréographe, etc.). Le décor représente un environnement de type "bureau" et est illustré sur la figure 7.2. Les séquences mettent en scène deux acteurs appelés personnages A et B. Au total, neuf séquences audio-visuelles ont été captées dans ce décor, avec deux cadrages différents. Ces deux positions de prise de vue permettent de présenter les séquences dans deux boîtes scéniques : une boîte dite "étendue", où les objets visuels sont placés entre 2 et 7 m, et une boîte "restreinte" où les objets sont répartis entre 2 et 5 m. Ces deux configurations sont illustrées sur la figure 7.3. Il est à noter que dans les deux cas les objets visuels sont placés devant et derrière le plan de convergence des caméras (fixé à 3 m). Ainsi, lors de la diffusion des séquences 3D, les objets visuels seront restitués devant et derrière l'écran de projection.

FIGURE 7.2 – *Exemple de séquence captée : deux personnages discutent dans un bureau (personnage A, à droite, et B, à gauche).*

Séquences utilisées pour l'évaluation de la QoE audio-visuelle 3D

Les neuf séquences captées sont décrites dans le tableau 7.1 et mettent en scène un ou deux personnages dans différentes situations. Suivant les séquences, les personnages sont placés à différentes positions afin d'explorer l'effet du rendu de la distance des objets sonores suivant plusieurs facteurs. Cet effet peut, par exemple, dépendre de la position absolue des locuteurs (devant ou derrière l'écran), ou encore de l'étendue en distance de la zone d'intérêt. La nature des éléments sonores présentés varie suivant les séquences, avec des contenus de type voix, musique, ou bruit de clavier.

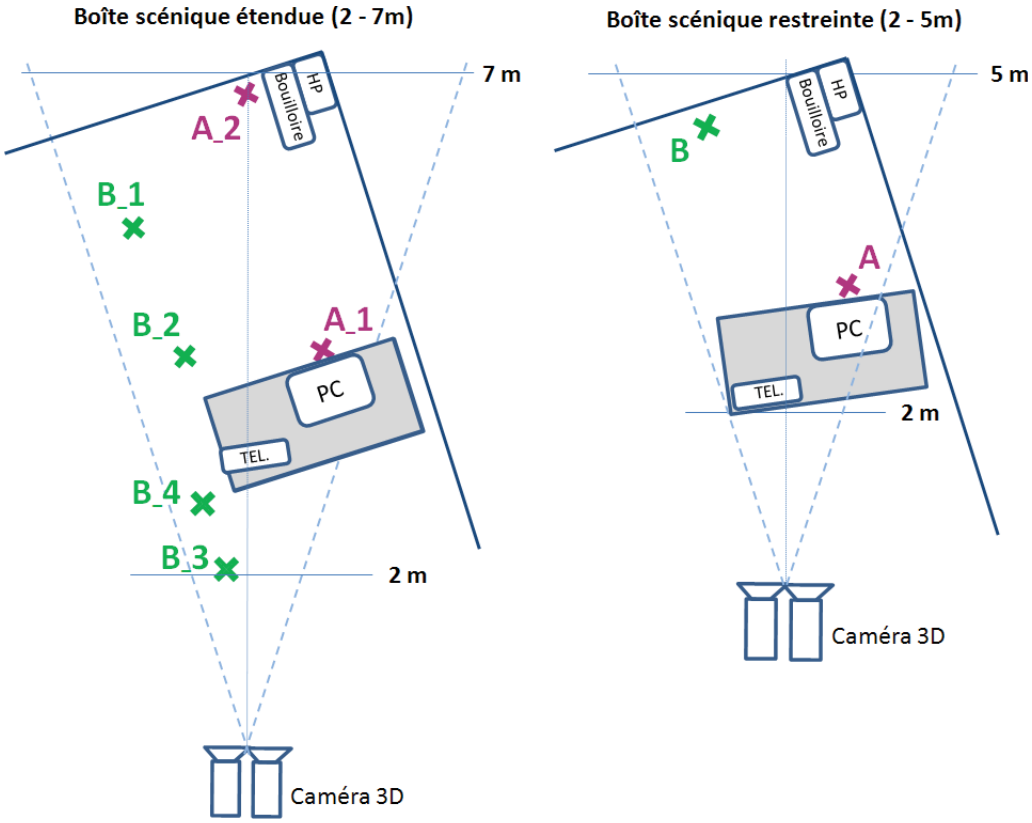


FIGURE 7.3 – *Illustration des éléments captés dans les deux boîtes scéniques, et des positions prises par les personnages A et B en fonction des séquences (voir tableau 7.1).*

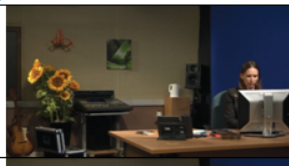
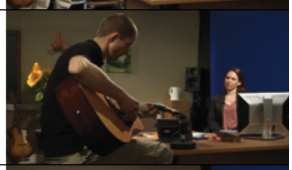
Séq.	Boîte scénique	*Contenu sonore*	Description	Position du personnage A (distance ; azimut)	*Position du personnage B (distance ; azimut)*	Capture d'écran
1	Etendue	Bruit de clavier	Personnage A (position 1) tape à l'ordinateur	4 m ; 12°	-	
2	Etendue	Voix	Dialogue entre les personnages A (position 1) et B (position 1)	4 m ; 10°	5.5 m ; -12°	
3	Etendue	Voix	Dialogue entre les personnages A (position 1) et B (position 2)	4 m ; 10°	4 m ; -11°	
4	Etendue	Musique	Personnage B (position 3) joue de la guitare	-	2 m ; -6°	
5	Etendue	Musique	Personnage B (position 4) joue de la guitare	-	2.6 m ; -10°	
6	Etendue	Voix	Personnage A (position 1) parle au téléphone	4 m ; 10°	-	
7	Etendue	Voix	Dialogue entre les personnages A (position 1) et B (position 4)	4 m ; 10°	2.6 m ; -10°	
8	Etendue	Voix	Dialogue entre les personnages A (position 2) et B (position 4)	7 m ; 0°	2.6 m ; -10°	
9	Restreinte	Voix	Dialogue entre les personnages A et B	3 m ; 10°	4.5 m ; -10°	

TABLE 7.1 – *Description de séquences audio-visuelles 3D captées.*

7.3 Évaluation de la QoE audio-visuelle 3D avec ou sans restitution sonore suivant la distance

7.3.1 Motivations

Nous souhaitons, à travers cette expérience subjective, évaluer l'apport potentiel du rendu de la distance d'une scène sonore lors de la visualisation de contenus audio-visuels 3D. Les expériences précédentes laissent en effet penser que la présentation conjointe d'éléments sonores suivant la distance et d'images en relief pourrait améliorer la qualité d'expérience audio-visuelle.

Cette expérience a également pour but de mesurer l'impact potentiel de la restitution du "relief sonore" sur la perception du relief visuel. En effet, même si aucun effet ventriloque inverse n'a pu être observé dans la section 6.2, nous pensons que la sensation de profondeur visuelle induite par les images 3D peut potentiellement être modifiée par la présence de stimuli sonores suivant la distance. Notons que si cette hypothèse est vérifiée, la restitution de la distance sonore représenterait une piste pour améliorer le confort de visualisation de contenus 3D (voir section 6.2.1).

Enfin, il est important, dans cette expérience, de mesurer l'impact de la restitution sonore en distance par WFS sur la qualité de l'expérience sonore des spectateurs. En effet, il est tout à fait possible que la restitution sonore testée ait une incidence uniquement sur la modalité auditive sans modifier l'expérience visuelle ou audio-visuelle d'un point de vue plus global.

7.3.2 Description des séquences audio-visuelles 3D évaluées

Les séquences audio-visuelles utilisées dans cette expérience subjective sont basées sur les neuf séquences décrites dans le tableau 7.1. Afin d'évaluer l'apport potentiel de la restitution sonore en distance, nous souhaitons présenter ces neuf séquences en présence ou non d'éléments sonores synthétisés suivant la distance. Pour cela, trois mixages de la scène sonore sont réalisés pour accompagner les séquences visuelles. Ces trois mixages sont appelés "sans distance", "distance réaliste" et "distance augmentée" (voir figure 7.4).

Dans le mixage "sans distance", les éléments sonores sont tous restitués au niveau de l'écran (à 3 m), en tenant compte de l'azimut des objets visuels à l'origine des sons émis. Il est à noter que la disparité en profondeur entre l'élément visuel et l'objet sonore (toujours placé à 3 m) est donc variable suivant les séquences.

Dans le mixage "distance réaliste", les éléments sonores sont synthétisés à la position des objets visuels présentés en relief (azimut et distance). Ainsi, si un objet visuel est placé à 1 m devant l'écran (en jaillissement), l'objet sonore correspondant est également synthétisé à 1 m devant l'écran.

Enfin, dans le mixage "distance augmentée", les objets sonores sont restitués suivant le même azimut, mais avec des informations de distance amplifiées par rapport à la position des objets visuels. Un facteur multiplicateur différent est appliqué suivant que l'objet est placé devant ou derrière le dispositif AV3D. Ainsi, pour un objet visuel placé derrière l'écran, le son sera placé 1,5 fois plus loin que l'objet visuel, alors que dans le cas d'un objet visuel en jaillissement, le son sera restitué 1,3 fois plus près. Ce jeu de facteurs semble être le meilleur

compromis pour maximiser l'effet de relief sonore, tout en restant dans les limites des zones d'intégration déterminées dans la section 6.3. Le respect de ces limites permet d'offrir aux spectateurs une présentation d'objets audio-visuels spatialement cohérents.

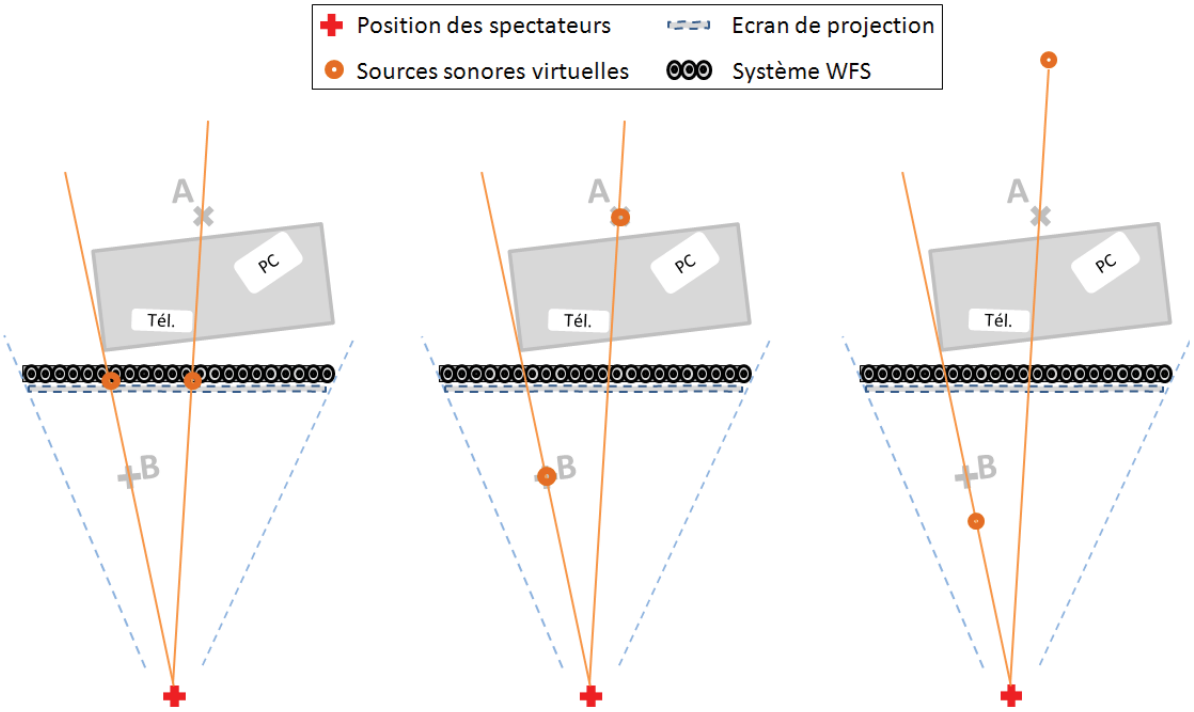


FIGURE 7.4 – *Illustration des positions de sources dans les trois mixages sonores proposés pour accompagner les vidéos 3D : "sans distance" (à gauche), "distance réaliste" (au milieu) et "distance augmentée" (à droite). Les éléments gris représentent les élément visuels perçus de part et d'autre du dispositif AV3D (personnages A et B, bureau, etc.).*

Les positions des sources sonores correspondant aux personnages A et B sont répertoriées dans le tableau 7.2. Un code couleur indique si les sources sonores sont placées à l'intérieur (en vert), en limite (en orange), ou à l'extérieur (en rouge) des zones d'intégrations audio-visuelles estimées grâce aux résultats de la section 6.3 [2]. Ce tableau montre que des conditions spatialement non-cohérentes sont présentées avec le mixage sonore "sans distance".

2. Le code couleur est donné à titre informatif sur la base d'une extrapolation des résultats présentés dans la section 6.3. En effet, afin de déterminer avec certitude si les objets sonores sont placés dans les zones d'intégration, il faudrait évaluer ces fenêtres pour toutes les distances des objets visuels présents dans la scène, et avec les stimuli audio-visuels utilisés dans cette nouvelle expérience.

Séq.	*Mixage sonore « sans distance »*		*Mixage sonore « distance réaliste »*		*Mixage sonore « distance augmentée »*	
	Distance perso. A	*Distance perso. B*	*Distance perso. A*	*Distance perso. B*	*Distance perso. A*	*Distance perso. B*
1	3 m	-	4 m	-	6 m	-
2	3 m	3 m	4 m	5.5 m	6 m	8.2 m
3	3 m	3 m	4 m	4 m	6 m	6 m
4	-	3 m	-	2 m	-	1.55 m
5	-	3 m	-	2.6 m	-	2 m
6	3 m	-	4 m	-	6 m	-
7	3 m	3 m	4 m	2.6 m	6 m	2 m
8	3 m	3 m	7 m	2.6 m	10.5 m	2 m
9	3 m	3 m	3 m	4.5 m	3 m	6.8 m

TABLE 7.2 – *Distance des sources sonores synthétisées dans les trois mixages sonores : "sans distance", "distance réaliste" et "distance augmentée". Le code couleur indique si les objets sonores sont placés à l'intérieur (en vert), en limite (en orange), ou à l'extérieur (en rouge) des zones d'intégrations audio-visuelles estimées par rapport aux positions des objets visuels (ces dernières étant les positions reportées pour le mixage "distance réaliste").*

7.3.3 Critères d'évaluation

Comme discuté dans la section 3.2.4, il peut être intéressant d'évaluer la qualité d'expérience liée à un service multimédia tel que la diffusion de contenus audio-visuels 3D grâce à plusieurs critères de notation. En effet, une évaluation multi-critères permet de qualifier la qualité des médias audio et vidéo, mais aussi la qualité de l'expérience faite par les utilisateurs en termes de confort, d'immersion, etc. Pour cette raison, les participants doivent juger les séquences audio-visuelles 3D suivant cinq critères de notation, explicités sous forme de questions :

- la **profondeur visuelle** : avez-vous l'impression que les éléments visuels sont répartis en profondeur (image en "relief"), et non au niveau de l'écran uniquement (image "plate") ?
- la **gêne visuelle** : êtes-vous gêné par la vidéo 3D (images doubles, flou, difficulté à faire la mise au point, sensation de 3D "agressive", etc.) ?
- la **profondeur sonore** : avez-vous l'impression que les éléments sonores sont répartis en profondeur (sensation de "relief sonore"), et non uniquement au niveau de l'écran ?
- la **qualité sonore** : est-ce que le son est de bonne qualité (sans sons distordus ni bruits parasites, etc.) ?
- l'**immersion audio-visuelle** : avez-vous l'impression d'être présent, immergé dans la scène ? La scène vous semble-t-elle réelle (cohérence entre son et image, etc.) ?

Pour effectuer leur jugement, les participants utilisent une échelle de notation discrète allant de 0 à 10 pour chaque critère. La note 0 indique que la réponse à la question posée est "non, pas du tout", alors que la note 10 correspond à la réponse "oui, beaucoup".

7.3.4 Protocole expérimental

Le test se déroule en trois phases. Dans un premier temps, une étape de familiarisation au contenu est proposée, dans laquelle un échantillon d'extraits est présenté. Lors de cette phase de découverte, le participant observe les conditions sans avoir à les noter. Ensuite, le participant se voit proposer une phase d'entraînement à la tâche, lors de laquelle il attribue des notes aux séquences sélectionnées, pour les cinq critères. Enfin, lors de la phase de test, l'intégralité des 27 extraits (neuf séquences visuelles accompagnées par trois mixages sonores) est présentée dans un ordre aléatoire et évaluée par le spectateur. Il faut préciser que le participant peut visualiser chaque extrait autant de fois que nécessaire avant d'attribuer les cinq notes. Lorsque tous les critères sont évalués, le participant passe à l'extrait suivant.

L'expérience se déroule dans la salle de test utilisée dans les expériences précédentes (superficie d'environ 20 m^2, niveau de bruit de fond de 35 $dB(A)$, $Tr_{60} \simeq 350\ ms$). Les contenus audio-visuels sont diffusés grâce au système de restitution AV3D, et les participants sont installés à une distance de 3 m de l'écran.

Le test est mené sur un panel constitué de 24 participants (13 femmes et 11 hommes), dont la moyenne d'âge est d'environ 31 ans. La majorité des sujets a participé à un ou plusieurs tests portant sur la perception de la distance d'objets unimodaux ou bimodaux (expériences décrites dans les chapitres précédents). Ces personnes sont donc sensibilisées aux tâches de localisation d'objets virtuels suivant la distance, et à la visualisation d'objets en 3D.

7.3.5 Résultats

Les notes des 24 participants sont soumises à une analyse de la variance de type ANOVA pour chacun des cinq critères d'évaluation : la profondeur visuelle, la gêne visuelle, la profondeur sonore, la qualité sonore et l'immersion audio-visuelle. Les résultats de ces analyses sont présentés ci-après.

Critères liés à l'expérience sonore

Les participants ont estimé la **profondeur sonore** de chaque extrait audio-visuel. Les résultats de l'analyse ANOVA montrent que les participants ont détecté les différences entre les trois mixages sonores en termes de profondeur sonore restituée ($F(2,46)=13.99$, $p<0.001$). Les jugements de profondeur sonore dépendent également de la séquence évaluée ($F(8,184)=6.12$, $p<0.001$). La figure 7.5 présente les notes moyennes et intervalles de confiance à 95 % associés en fonction des séquences.

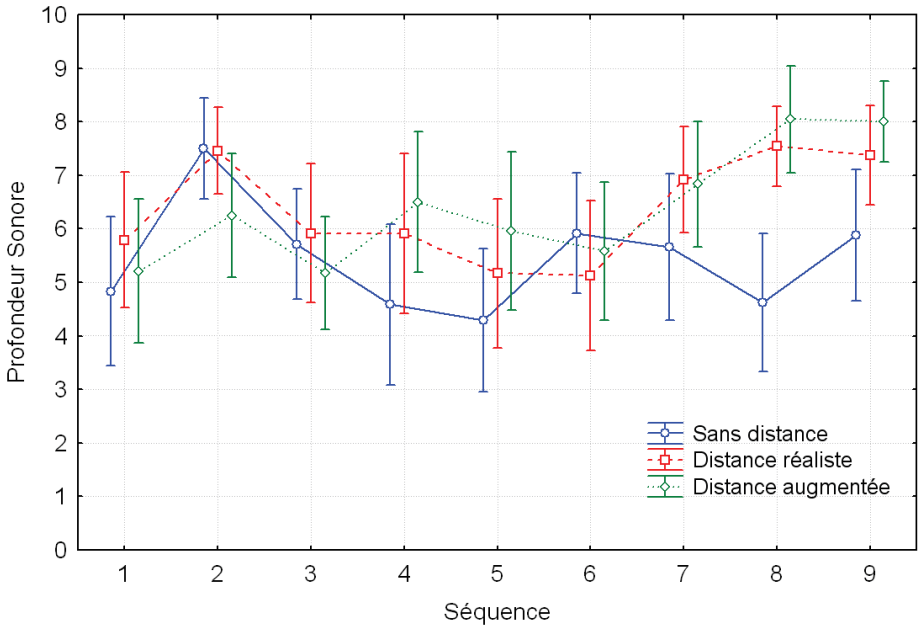


FIGURE 7.5 – *Notes de profondeur sonore estimées pour les neuf séquences accompagnées par les trois mixages sonores : "sans distance" (en bleu), "distance réaliste" (en rouge) et "distance augmentée" (en vert).*

La figure 7.5 montre que les notes de profondeur sonore sont généralement supérieures pour les mixages avec restitution de la distance que pour le mixage "sans distance". Cet effet est particulièrement marqué pour certaines séquences (4, 8 et 9 par exemple). Notons cependant que pour les séquences 1, 3 et 6, aucune différence significative n'est détectée en termes de profondeur sonore entre les trois mixages sonores présentés. En se référant au tableau 7.1, on constate que les séquences 1 et 6 mettent en scène un seul personnage qui est placé derrière l'écran. La séquence 3 présente quant à elle deux personnages placés à la même distance (4 m, soit 1 m derrière l'écran). Pour ces trois séquences, les différences entre les mixages entraînent donc une variation du niveau sonore en absolu sans créer de différence entre les sources présentes dans la scène. Or, la définition de la profondeur sonore qui est fournie aux participants fait référence à la répartition des éléments sonores présents dans la scène suivant la profondeur (voir section 7.3.3). Conformément à cette définition, il est probable que les sujets n'interprètent pas un changement de niveau sonore absolu comme une différence de profondeur. Notons que cette hypothèse ne permet pas d'expliquer les notes attribuées pour la séquence 2. De plus, les séquences 4 et 5 ne présentent qu'un seul élément sonore, mais ont été notées différemment suivant les mixages sonores. Il semble donc que cette hypothèse ne permette pas à elle seule d'expliquer les différences de profondeur sonore perçues entre les mixages.

La **qualité sonore** constitue le deuxième critère d'évaluation relatif à l'expérience sonore. L'analyse ANOVA révèle que la qualité sonore perçue est comparable pour les trois mixages audio présentés ($F(2,46)=2.63$, $p>0.05$). Les jugements de qualité dépendent cependant des séquences ($F(8,184)=6.74$, $p<0.001$) avec un effet d'interaction notable entre le mixage et

les séquences (F(16,368)=4.38, $p<0.001$). La figure 7.6 illustre les notes moyennes de qualité attribuées par les participants et les intervalles de confiance à 95 % associés.

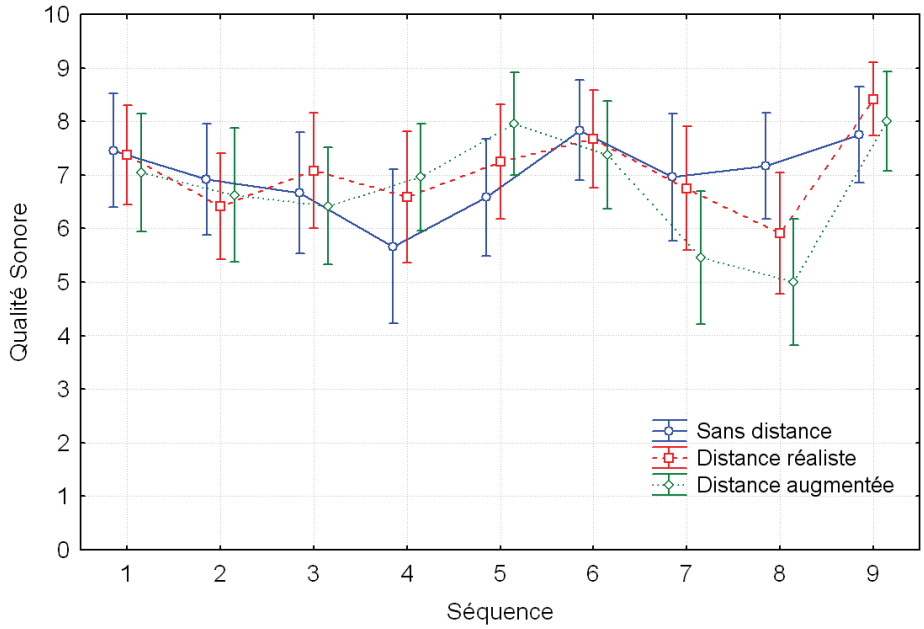


FIGURE 7.6 – *Notes de qualité sonore et intervalles de confiance à 95 % associés pour les trois mixages sonores : "sans distance" (en bleu), "distance réaliste" (en rouge) et "distance augmentée" (en vert).*

Nous pouvons observer sur la figure 7.6 que la qualité perçue des trois mixages audio est comparable pour la plupart des séquences. Cependant, des différences apparaissent pour les séquences 4 et 5, où les mixages "distance réaliste" et "distance augmentée" sont mieux notés que le mixage "sans distance". Au contraire, le mixage "distance augmentée" est perçu comme étant de moins bonne qualité pour les séquences 7 et 8. Notons que, dans ce dernier cas, plusieurs hypothèses peuvent être avancées pour expliquer la dégradation de qualité perçue.

Pour commencer, les séquences 7 et 8 sont celles qui présentent le plus grand écart en distance entre les éléments sonores. En effet, si on se réfère au tableau 7.2, la différence de distance entre les personnages A et B atteint 4 m pour la séquence 7 et 8, 5 m pour la séquence 8. Or, plus l'écart entre les personnages est important, plus la différence de niveau sonore entre les locuteurs est marquée lors de la diffusion. Il est possible que cette différence de niveau sonore (présente au sein d'un même extrait) affecte l'intelligibilité de la scène sonore, et dégrade la qualité audio perçue par les spectateurs.

Ensuite, il est également probable que les spectateurs aient perçu une dégradation spectrale dans les séquences 7 et 8. Une première hypothèse allant dans ce sens est basée sur une limitation de la Wave Field Synthesis. En effet, il a été montré que des phénomènes de coloration spectrale peuvent apparaître lors de la restitution de sources acoustiques focalisées [Spors et al., 2009, Wierstorf et al., 2013]. Ces phénomènes sont liés aux approximations nécessaires au passage de la théorie de la WFS à sa mise en œuvre pratique sur un réseau de haut-parleurs

(voir section 2.2.5). Une autre hypothèse repose sur un défaut de réalisme du spectre de la voix du locuteur. En effet, dans les séquences 7 et 8, le personnage qui parle est placé de dos (voir captures d'écran du tableau 7.1). Dans une situation réelle, les composantes hautes fréquences de la voix du locuteur devraient être atténuées à cause de leur directivité et des phénomènes de diffraction acoustique. Or, dans notre cas, la prise de son a été réalisée au moyen d'un microphone cravate, placé au niveau du col de l'acteur. En d'autres termes, le positionnement du microphone ne tient pas compte de l'orientation du personnage par rapport au spectateur. Il est possible que les spectateurs soient capables de détecter une incohérence entre l'orientation du personnage et le timbre de voix restitué, ce dernier étant anormalement riche en hautes fréquences. Cette incohérence serait d'autant plus importante dans le cas des mixages "distance réaliste" et "distance augmentée", où le niveau sonore de la voix du personnage est plus élevé.

Critères liés à l'expérience visuelle

Les participants estiment également la **profondeur visuelle** des 27 extraits diffusés. Rappelons que dans les trois versions présentées de chaque séquence, seules les composantes sonores sont modifiées. Les éléments visuels sont, pour leur part, strictement identiques d'une version à l'autre. Les notes moyennes de profondeur visuelle et intervalles de confiance à 95 % sont reportés sur la figure 7.7.

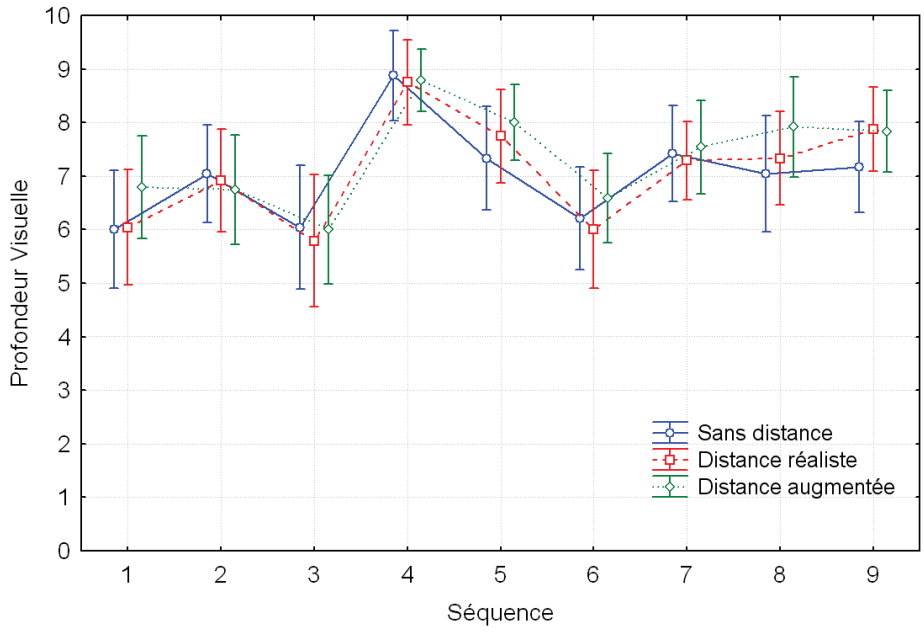


FIGURE 7.7 – *Notes de profondeur visuelle estimées pour les neuf séquences accompagnées par les trois mixages sonores : "sans distance" (en bleu), "distance réaliste" (en rouge) et "distance augmentée" (en vert).*

Nous pouvons observer sur la figure 7.7 que les jugements de profondeur visuelle dépendent fortement des séquences, avec des notes comprises entre 6 et 9, respectivement pour les séquences 3 et 4. Cet effet de la séquence est confirmé par l'analyse ANOVA ($F(8,184)=10.85$,

p<0.001). Il est intéressant de voir que l'étendue de la boîte scénique ne semble pas impacter la perception de la profondeur visuelle. En effet, la séquence 9, pour laquelle la boîte scénique est restreinte (objets placés entre 2 et 5 m), semble offrir la même profondeur visuelle que les séquences captées avec la boîte scénique étendue (séquences 7 ou 8, par exemple). Les notes obtenues pour la séquence 9 sont même supérieures à certaines séquences correspondant à la boîte scénique étendue (séquences 3 et 6, par exemple). Ces résultats indiquent que les jugements de profondeur visuelle ne tiennent pas uniquement compte des disparités affichées à l'écran. Il est probable que d'autres indices visuels, comme la perspective linéaire créée par le décor par exemple, influencent les jugements des spectateurs. En effet, les captures d'écran présentées dans le tableau 7.1 montrent que, pour la séquence 9, la perspective linéaire formée par les murs en arrière plan est différente des autres séquences.

Ensuite, l'analyse ANOVA révèle un léger effet du mixage sonore sur la perception de la profondeur visuelle ($F(2,46)=3.29$, *p<0.05*). La figure 7.7 permet en effet de constater que, pour les séquences 5, 8 et 9 notamment, le mixage "distance augmentée" obtient des notes légèrement supérieures aux autres mixages bien que le contenu visuel soit strictement identique. Notons que ces séquences correspondent aux extraits offrant le plus de différences entre les mixages sonores en termes de profondeur sonore perçue (voir figure 7.5) [3]. Ce résultat aurait tendance à confirmer l'hypothèse suivant laquelle la perception de la profondeur visuelle peut être biaisée par des informations sonores. Cependant, devant la faiblesse de cet effet, il reste difficile de tirer une conclusion sur ce point.

Concernant le critère de **gêne visuelle**, les jugements semblent uniquement dépendre de la séquence évaluée ($F(8,184)=3.88$, *p<0.001*), et non du mixage sonore ($F(2,46)=1.33$, *p=0.27*). Ces résultats étaient attendus puisque la gêne visuelle est principalement induite par la présence des disparités binoculaires, et que ce facteur varie en fonction des séquences, mais pas en fonction des conditions audio. La figure 7.8 présente les résultats de l'expérience subjective pour le critère lié à la gêne visuelle.

3. La figure 7.5 montre également une influence du mixage pour la séquence 4 qui n'est pas observée pour le critère de profondeur visuel (figure 7.7). Il est probable que la note particulièrement élevée de profondeur visuelle obtenue pour le mixage "sans distance" ($\simeq 0.9$) agisse comme une "valeur plafond" et limite ainsi l'influence potentielle de la profondeur sonore sur la profondeur visuelle.

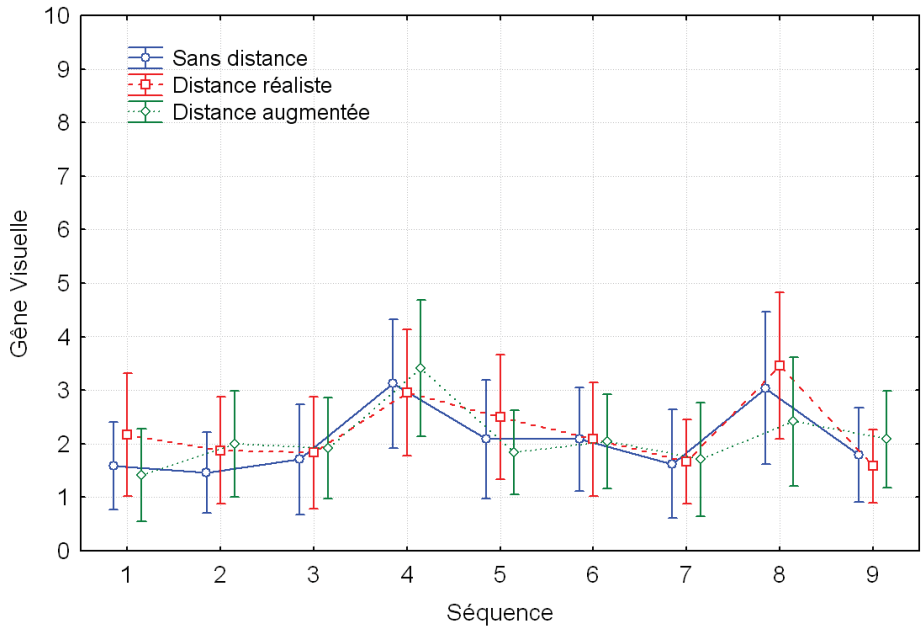


FIGURE 7.8 – *Notes de gêne visuelle et intervalles de confiance à 95 % associés pour les trois mixages sonores : "sans distance" (en bleu), "distance réaliste" (en rouge) et "distance augmentée" (en vert).*

Nous pouvons constater sur la figure 7.8 que la gêne occasionnée par la visualisation des séquences est relativement limitée, avec des notes inférieures à 3,5. Cependant, les séquences 4 et 8 semblent davantage critiques en termes de confort de visualisation. Ce résultat s'explique certainement par la position des éléments visuels de la scène. En effet, dans la séquence 4, le personnage B est présenté à une distance de 2 m (soit un jaillissement de 1 mètre devant l'écran). Cette position coïncide, à une dizaine de centimètres près, avec la limite de la zone de confort définie dans la section 4.3.2 ($1,87$ m pour la distance minimale, et $7,5$ m pour la distance maximale). Cette hypothèse est d'ailleurs confirmée par la diminution de la gêne visuelle observée pour la séquence 5, dans laquelle le personnage B est présenté avec un jaillissement modéré (personnage B placé à $2,6$ m au lieu de 2 m dans la séquence 4). La séquence 8 offre, quant à elle, la plus grande différence de position entre les objets visuels d'intérêt, avec les personnages A et B placés à des distances respectives de 7 et $2,6$ m (voir tableau 7.1). Bien que ces positions respectent les limites de la zone de confort, il est possible que les spectateurs soient gênés par le fait d'avoir à faire la mise au point sur les deux personnages successivement. Cette action de va-et-vient entre deux plans en profondeur nécessite en effet une adaptation rapide du système visuel en termes de distances de convergence et d'accommodation.

Critère lié à l'expérience audio-visuelle

Le dernier critère évalué dans cette étude concerne la sensation d'**immersion audio-visuelle**. Nous pensons que ce critère est particulièrement important pour l'évaluation de la qualité d'expérience audio-visuelle 3D. En effet, il constitue un concept de haut niveau qui peut potentiellement être impacté par les différents critères de notation évalués précédemment

(profondeur sonore, qualité sonore, profondeur visuelle, gêne visuelle). L'analyse ANOVA menée sur les résultats des 24 participants révèle que les notes d'immersion audio-visuelle sont à la fois influencées par la séquence présentée ($F(8,184)=2.88$, $p<0.005$) et par le mixage sonore ($F(2,46)=7.68$, $p=0.001$). La figure 7.9 illustre ces effets à travers les jugements moyens attribués par les 24 participants et les intervalles de confiance à 95 % associés pour ce critère.

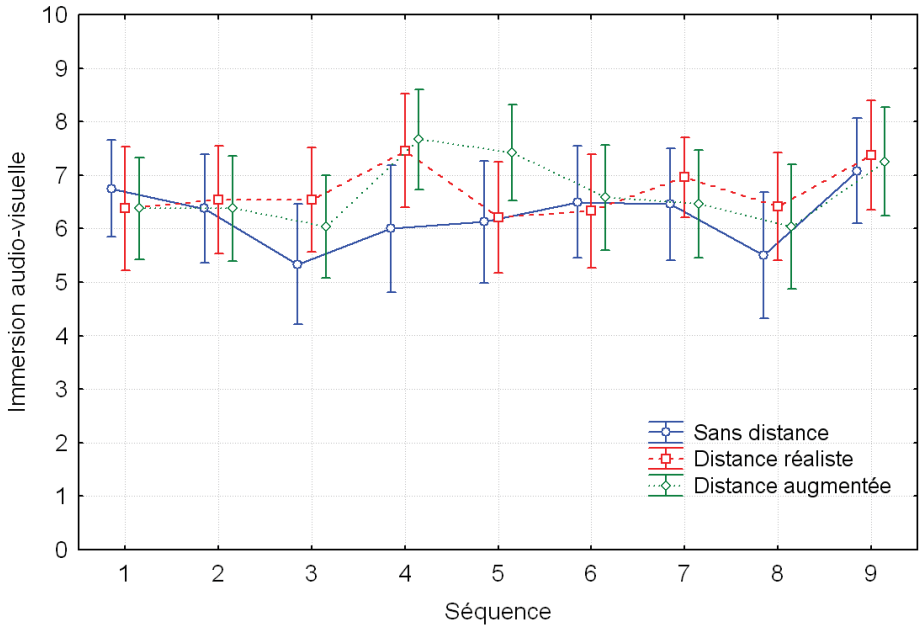


FIGURE 7.9 – *Notes d'immersion audio-visuelle et intervalles de confiance à 95 % associés pour les trois mixages sonores : "sans distance" (en bleu), "distance réaliste" (en rouge) et "distance augmentée" (en vert).*

Nous pouvons observer sur la figure 7.9 que, pour toutes les séquences évaluées, les mixages "distance réaliste" et "distance augmentée" obtiennent des notes d'immersion supérieures ou égales au mixage "sans distance". Les différences entre les mixages sonores apparaissent plus nettement pour les séquences 4 et 5. Il est à noter que ces deux séquences présentent un objet visuel en jaillissement associé à un stimulus sonore de type musical (voir tableau 7.1). L'hypothèse suivant laquelle le critère d'immersion audio-visuelle serait impacté par les quatre autres critères d'évaluation pourrait expliquer l'augmentation de la sensation d'immersion des séquences 4 et 5 pour le mixage "distance augmentée". En effet, les résultats précédents montrent que ce mixage offre davantage d'effets de profondeur sonore (voir figure 7.5), et une meilleure qualité sonore (figure 7.6). De plus, la profondeur visuelle semble légèrement plus importante dans le cas du mixage "distance augmentée" pour la séquence 5 (figure 7.7) sans pour autant dégrader le confort de visualisation (figure 7.8).

Nous souhaitons à présent obtenir des informations plus précises quant aux relations existantes entre les cinq critères d'évaluation utilisés lors de cette expérience. Pour cela, une analyse des corrélations est menée sur les cinq critères. Les résultats de cette analyse sont présentées dans le tableau 7.3.

Critère de notation	*Gêne visuelle*	*Profondeur visuelle*	*Profondeur sonore*	*Qualité sonore*	*Immersion AV*
Gêne visuelle	1	-0.24	-0.13	-0.37	-0.40
Profondeur visuelle	-	1	0.40	0.30	0.56
Profondeur sonore	-	-	1	0.42	0.55
Qualité sonore	-	-	-	1	0.59
Immersion AV	-	-	-	-	1

TABLE 7.3 – *Coefficients de corrélation calculés sur les cinq critères d'évaluation de la qualité d'expérience audio-visuelle 3D.*

Il apparaît, à la lecture du tableau 7.3, que l'immersion audio-visuelle obtient les coefficients de corrélation les plus élevés (en valeur absolue). Ce résultat indique que chaque critère (gêne visuelle, profondeur visuelle, profondeur sonore et qualité sonore) participe à la sensation d'immersion. Notons que les coefficients de corrélation négatifs obtenus pour le critère de gêne visuelle indiquent qu'un extrait est jugé comme étant moins immersif lorsque la gêne visuelle perçue augmente.

7.3.6 Discussions

Apport du rendu du "relief sonore" sur l'expérience audio-visuelle 3D

Les différents résultats de cette expérience subjective indiquent que l'apport du rendu du "relief sonore" sur l'expérience audio-visuelle 3D varie suivant la séquence. En effet, trois types de séquences se distinguent :

— celles pour lesquelles la profondeur sonore est détectée et contribue à une meilleure expérience audio-visuelle, notamment avec une augmentation des notes d'immersion (séquences 4 et 5, par exemple),
— celles pour lesquelles l'expérience audio-visuelle 3D reste inchangée en termes d'immersion, bien que des changements de profondeur sonore soient détectés (séquences 7 et 8, par exemple),
— enfin, les séquences pour lesquelles les différents mixages ne sont pas perçus en termes de profondeur sonore (séquences 3 et 6, par exemple).

La figure 7.10 illustre un exemple de séquence pour chacune des deux premières catégories, c'est-à-dire lorsque la profondeur sonore des différents mixages est détectée.

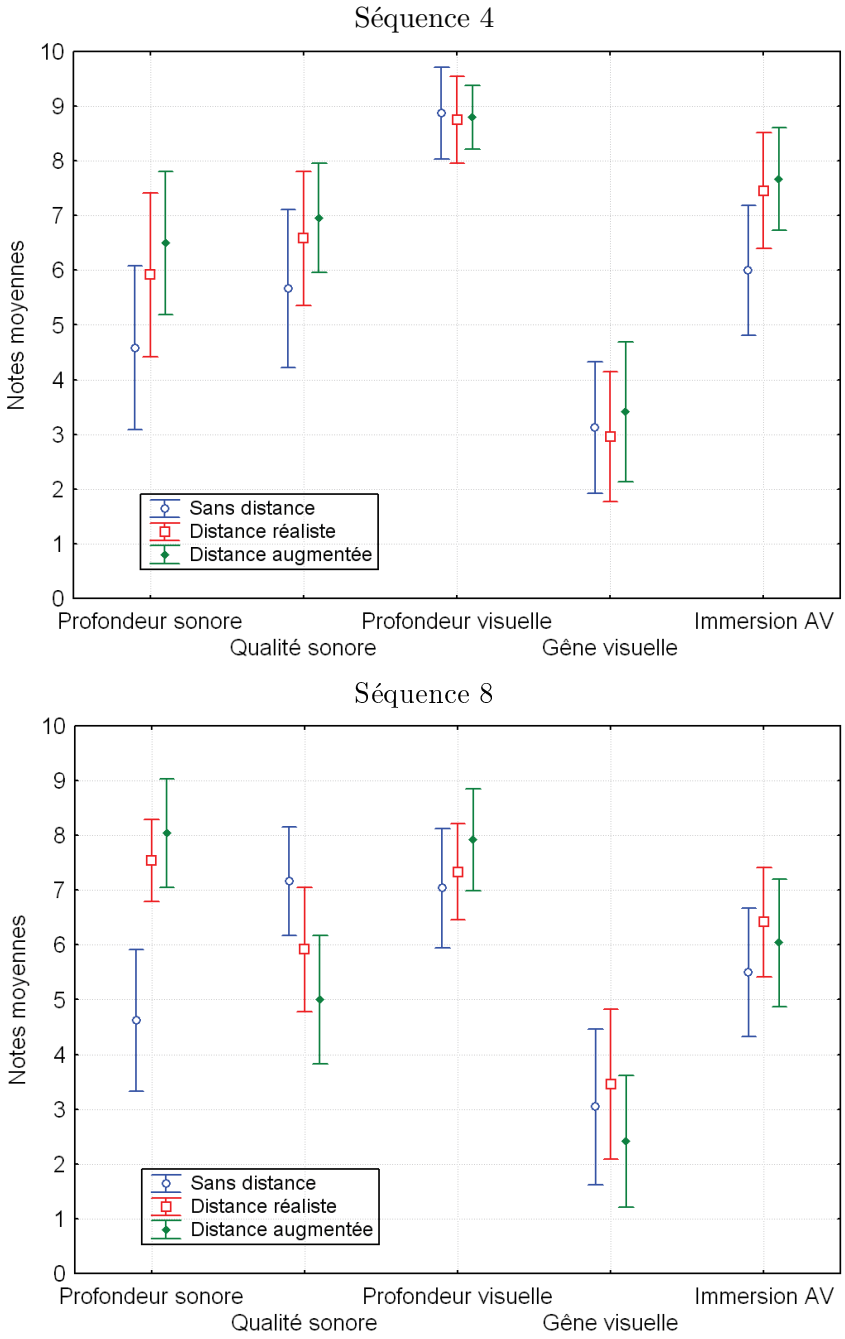


FIGURE 7.10 – *Notes obtenues pour tous les critères dans le cas des séquences 4 (en haut) et 8 (en bas).*

La figure 7.10 montre, pour la séquence 4, une sensibilité de certains critères aux mixages sonores présentés. Pour ce type de séquence, il est possible d'affirmer que le rendu du "relief sonore" améliore la qualité d'expérience dès lors qu'une augmentation des notes de profondeur sonore, de qualité sonore et surtout d'immersion audio-visuelle est observée. En revanche, la séquence 8 illustre un exemple pour lequel le rendu du "relief sonore" engendre une augmentation de la profondeur sonore (et visuelle) au détriment de la qualité sonore perçue. D'après

les relations de corrélation répertoriées dans le tableau 7.3, la perception d'une telle dégradation impacte la sensation d'immersion. Notons cependant que même si la note d'immersion est potentiellement altérée par une diminution de la qualité sonore dans le cas de la séquence 8, les mixages avec rendu de la distance semblent offrir un niveau d'immersion comparable aux autres mixages d'un point de vue statistique. Cette observation est d'ailleurs valable pour la plupart des séquences puisqu'il a été montré sur la figure 7.9 que les notes d'immersion des mixages sonores avec rendu de la distance sont supérieures ou égales à celles du mixage "sans distance".

Le rendu du "relief sonore" pour lutter contre la gêne visuelle ?

La comparaison des résultats obtenus pour les séquences 4 et 5 permet d'apporter des éléments de réponse quant à la question de l'amélioration de l'expérience visuelle 3D grâce à la restitution du "relief sonore" (figure 7.11). Ces deux séquences ont pour seule différence la proportion de jaillissement du personnage B, qui est placé à 2 m dans la séquence 4, et à $2,6$ m dans la séquence 5.

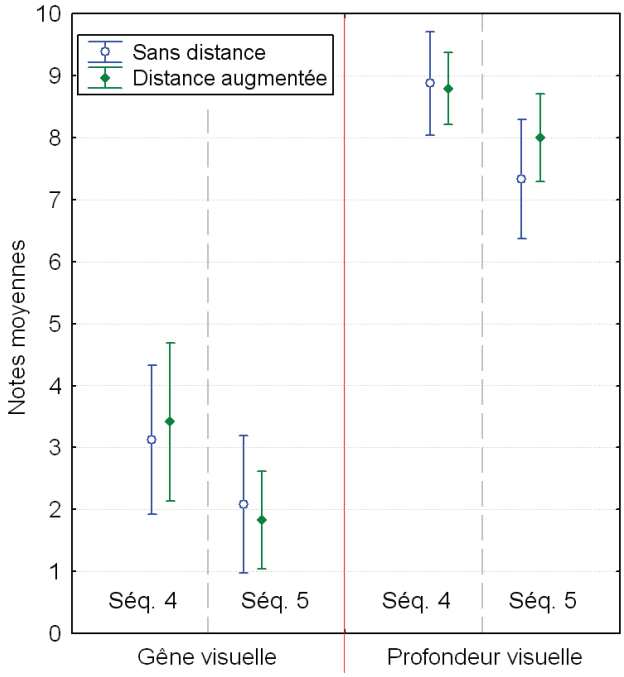


FIGURE 7.11 – *Notes de gêne visuelle et de profondeur visuelle pour les séquences 4 et 5 accompagnées des mixages "sans distance" (en bleu) et "distance augmentée" (en vert).*

Nous avons déjà mis en évidence que le changement de position du personnage B a pour conséquence une diminution de la gêne visuelle perçue dans la séquence 5. La figure 7.11 illustre également la diminution de la profondeur visuelle perçue entre les séquences 4 et 5. Bien que la condition "distance augmentée" obtienne une note légèrement supérieure au mixage "sans distance" pour la séquence 5, la profondeur visuelle perçue reste inférieure à celle de la séquence 4. Le croisement de ces résultats semble indiquer que le "relief sonore" ne permet

pas de diminuer la gêne visuelle tout en maintenant une sensation de profondeur visuelle à un niveau constant. En effet, la présentation du mixage sonore "distance augmentée" n'entraîne pas une augmentation suffisante de la sensation de profondeur visuelle pour compenser le changement de position des éléments visuels entre les séquences 4 et 5 (diminution de 60 *cm* du jaillissement associé au personnage B).

7.4 Conclusion

Le but de ce chapitre était d'étudier l'impact du rendu de la distance d'objets sonores sur la perception de scènes audio-visuelles 3D. Pour cela, la qualité d'expérience audio-visuelle 3D a été évaluée par le biais d'une expérience subjective utilisant cinq critères de notation (profondeur sonore, qualité sonore, profondeur visuelle, gêne visuelle, immersion audio-visuelle). Les stimuli présentés aux spectateurs ont été spécialement créés pour les besoins de l'expérience et tiennent compte des contraintes liées au système de restitution AV3D utilisé. Les participants ont ainsi évalué neuf séquences 3D présentées avec trois mixages sonores où la distance des objets sonores variait.

D'une manière générale, les jugements récoltés suivant les différents critères d'évaluation ont montré que la restitution de la distance des objets sonores augmente la profondeur perçue de la scène sonore sans impacter la qualité de la restitution sonore. De plus, la restitution du "relief sonore" semblerait augmenter la perception de la profondeur visuelle de manière sensible (sans engendrer de gêne visuelle). Cet élément nécessiterait d'être étudié de manière plus approfondie, car la tendance observée reste faible. Enfin, le mixage sonore a également un impact positif sur l'immersion audio-visuelle, puisque les extraits sont généralement jugés plus immersifs lorsque les séquences visuelles sont présentées avec les mixages "distance réaliste" ou "distance augmentée". Cependant, les résultats de cette expérience ont également révélé que les jugements dépendent de la séquence présentée. Pour certaines séquences, par exemple, la restitution de la distance des objets sonores ne semble pas être détectée par les spectateurs. Dans ce cas, les notes attribuées aux autres critères sont relativement stables en fonction du mixage sonore. Pour d'autres séquences, la restitution de la distance peut entraîner une dégradation de la qualité sonore perçue.

Ces résultats indiquent que la restitution de scènes sonores et visuelles spatialement cohérentes peut, dans certains cas, améliorer la qualité d'expérience audio-visuelle 3D en augmentant notamment la sensation d'immersion. L'apport du rendu de la distance des objets sonores est d'autant plus marqué pour les séquences mettant en scène des objets visuels en jaillissement avec des contenus musicaux (séquences 4 et 5). Il y a fort à penser que cette observation aurait été valable pour toutes les séquences présentant des éléments visuels en jaillissement (séquences 4, 5, 7 et 8) si la restitution du "relief sonore" n'avait pas engendré une dégradation de la qualité sonore perçue pour les séquences 7 et 8. L'utilisation d'une loi d'atténuation du niveau sonore plus modérée [4] aurait peut-être limité la dégradation de la qualité sonore perçue pour ces deux séquences.

4. Loi d'atténuation de l'ordre de 4 à 5 dB par doublement de distance par exemple, au lieu de 6 dB comme utilisé dans cette expérience.

Conclusion générale

L'objectif principal de ce travail de thèse était d'évaluer s'il est possible d'améliorer la qualité d'expérience audio-visuelle 3D grâce à la restitution sonore, et si oui, dans quelle mesure. Pour répondre à cette problématique, nous avons réalisé des expérimentations sur la thématique de la perception audio-visuelle. Ces expériences ont été menées autour de trois axes de recherche : la perception de la distance d'objets virtuels, la perception de la cohérence spatiale audio-visuelle et l'évaluation de la qualité d'expérience audio-visuelle 3D. Les principales contributions de ce travail et les perspectives qui peuvent en être dégagées sont détaillées dans les paragraphes suivants.

Contributions de la thèse

Nous avons essayé, tout au long de ce travail, de répondre aux différentes questions énoncées dans l'introduction générale de ce document. Nous proposons de nous appuyer sur ces questions pour illustrer les contributions principales de ce travail.

- **Comment l'expérience audio-visuelle 3D proposée grâce aux technologies disponibles actuellement est-elle perçue ? Est-il nécessaire d'introduire une nouvelle technologie de restitution sonore pour accompagner les vidéos 3D ?** Nous avons proposé, dans le chapitre 3, un protocole expérimental permettant d'évaluer l'expérience audio-visuelle 3D telle qu'elle peut être perçue aujourd'hui par le grand public, au cinéma ou à la télévision, par exemple. Ce protocole repose sur une approche multi-critères qui permet de rendre compte de l'expérience sonore, visuelle et audio-visuelle. Différents systèmes de restitution sonore couramment utilisés dans les applications destinées au grand public ont été testés pour accompagner les images 3D : un casque audio, une barre de son et un système multicanal 5.1. Cette étude a révélé qu'aucun de ces systèmes ne semble plus adapté que les autres pour accompagner des vidéos stéréoscopiques. De plus, cette expérience a montré que la sensation d'immersion audio-visuelle est fortement corrélée avec la notion de cohérence perçue entre le son et l'image. Compte tenu de ces observations, nous avons eu pour idée de proposer aux spectateurs des espaces audio-visuels offrant une meilleure cohérence, à l'aide d'un système de diffusion capable de restituer la distance d'objets sonores. Nous pensons en effet que la restitution de "relief sonore" est une piste prometteuse pour accompagner le relief des images stéréoscopiques.

- **Est-il possible de proposer aux spectateurs une restitution sonore offrant une sensation de distance ? Si oui, grâce à quelle(s) technologie(s) ?**
  Nous avons décrit, dans le chapitre 2, différents systèmes de restitution sonore potentiellement capables d'offrir des effets de relief sonore. Compte tenu du contexte applicatif visé (cinéma, télévision), nous avons mis en place un système associant la Wave Field Synthesis à un dispositif vidéo 3D. La réunion de ces technologies constitue un système de restitution audio-visuelle appelé AV3D (voir chapitre 4). Les capacités de la Wave Field Synthesis à reproduire des objets sonores perçus comme étant placés à différentes distances sont démontrées dans la section 5.3.1.

- **Quelle est la sensibilité des spectateurs à la restitution de la distance d'objets sonores par rapport au relief d'objets visuels ?**
  Les expériences menées dans le chapitre 5 ont montré que le système AV3D est capable de restituer des objets audio et/ou visuels suivant la distance, perçus comme étant placés devant et derrière le dispositif de diffusion. La distance des objets virtuels simulés (sonores, visuels ou audio-visuels) est généralement sous-estimée par les spectateurs. Nous avons également observé un phénomène de compression plus ou moins marqué suivant la modalité impliquée. Les jugements de distance d'objets sonores sont plus compressés et plus variables que ceux d'objets visuels ou audio-visuels.

- **Dans quelles limites des objets virtuels audio-visuels sont-ils perçus comme étant spatialement cohérents ?**
  Les expériences menées dans le chapitre 6 portent sur la perception audio-visuelle dans le cas d'objets sonores et visuels présentant des disparités suivant la distance. Sous certaines conditions, ces disparités semblent tolérées par notre système perceptif grâce à la flexibilité de notre cerveau. Les écarts pour lesquels les stimuli sont perçus comme un stimulus audio-visuel unique sont évalués dans la section 6.3. Ces écarts ne doivent pas dépasser un certain seuil afin de maintenir la sensation de cohérence audio-visuelle. Ces seuils de tolérance définissent les zones d'intégration spatiale audio-visuelle. On observe que l'étendue des zones d'intégration augmente avec la distance des objets visuels.

- **Est-ce que la restitution d'objets audio-visuels spatialement cohérents suivant la distance peut améliorer la qualité d'expérience des utilisateurs ?**
  Les conditions de cohérence spatiale définies dans le chapitre 5 sont exploitées dans le chapitre 7 par le biais d'un test subjectif. Trois mixages sonores sont alors présentés pour accompagner des séquences vidéos 3D. Un des mixages est dépourvu de restitution sonore suivant la distance alors que les deux autres mixages offrent des effets de "relief sonore" plus ou moins prononcés. Les résultats de cette expérience montrent que la restitution de la distance des objets sonores peut augmenter la profondeur perçue de la scène sonore et la sensation d'immersion audio-visuelle, sans dégrader la qualité de la restitution sonore. Cependant, ces résultats sont observés uniquement pour certaines séquences. Il est à noter que, dans certains cas, la restitution de la distance des objets sonores n'est pas clairement détectée par les spectateurs ou engendre une dégradation

de la qualité sonore perçue.

Perspectives dégagées

Ces travaux ont porté sur la capacité d'un système Wave Field Synthesis à fournir des informations sonores spatialement cohérentes avec une scène visuelle présentée en 3D. Cette étude dégage de nombreuses perspectives et fait émerger de nouvelles questions.

Pour commencer, notons que nous avons privilégié la Wave Field Synthesis car cette technologie semble particulièrement adaptée pour les applications visées (cinéma, télévision). En effet, ces dernières impliquent généralement la présence de plusieurs utilisateurs, placés au sein d'une zone d'écoute relativement étendue. L'absence de *sweet spot* et la restitution de l'effet de parallaxe acoustique font de la Wave Field Synthesis une solution technologique adaptée à ce type d'applications. Cependant, d'autres technologies de spatialisation sonore sont aussi capables de restituer des informations relatives à la distance. Le choix de la technologie de spatialisation est fortement conditionné par le contexte d'application visé. La technologie binaurale, par exemple, pourrait constituer une solution de spatialisation sonore adaptée pour un service de diffusion audio-visuelle destiné à un utilisateur seul, en mobilité.

Il serait alors intéressant d'étudier dans quelle mesure d'autres technologies de spatialisation sonore (binaural, ou Higher Order Ambisonic, par exemple) sont capables de restituer la distance d'un objet sonore virtuel.

Dans ce travail, la restitution sonore assurée par Wave Field Synthesis permet de reconstruire l'onde directe émise par des sources acoustiques virtuelles. La modification de la position en profondeur de ces dernières implique notamment des variations du niveau sonore. Ces variations contribuent à la sensation de "relief sonore" perçue par les spectateurs. Or, il a été montré dans le chapitre 2 qu'il existe d'autres indices perceptifs permettant d'estimer la distance d'un objet sonore. Il serait intéressant d'étudier les phénomènes d'interaction audio-visuelle 3D dans le cas d'un environnement sonore plus riche, en ajoutant des effets de réverbération, par exemple. Ces derniers pourraient ainsi rendre compte des variations du rapport "champ direct sur champ réverbéré" et créer une scène sonore virtuelle plus réaliste.

Ensuite, nous avons exploré, tout au long de cette étude, l'existence potentielle d'un effet ventriloque inverse. Nous souhaitions savoir si la présentation d'informations sonores suivant la distance pouvait biaiser la perception du relief d'un objet visuel présenté en 3D. Pour ce faire, nous avons choisi d'étudier ce phénomène suivant deux approches : via une tâche de localisation en distance (chapitre 6), mais aussi via l'évaluation du degré de profondeur visuelle d'une scène réaliste (chapitre 7).

Il n'a pas été possible, dans ces deux expériences, de prouver l'existence d'un effet ventriloque inverse. Cependant, bien qu'ils soient trop faibles pour permettre une conclusion, des effets statistiquement significatifs ont été observés. Il conviendrait donc de mener des études approfondies sur ce phénomène d'interaction audio-visuelle. Différentes hypothèses et pistes de travail ont d'ailleurs été énoncées dans le chapitre 6 (modification de la méthode de recueil

des jugements, des consignes de test, ou de la saillance des informations visuelles, par exemple).

Il a été montré, dans le chapitre 6, que le cerveau humain a la capacité d'intégrer des informations unimodales spatialement incohérentes en termes de distance (voir section 6.3). Il est important de souligner que, conjointement à la distance, d'autres dimensions de l'espace (azimut et élévation) pourraient être concernées par des écarts de localisation entre des stimuli visuels et sonores. Il serait donc pertinent d'étudier l'étendue des fenêtres d'intégration audio-visuelle en présence de disparités multi-dimensionnnelles. Ces dernières pourraient également porter sur des décalages temporels entre les informations sonores et visuelles.

Les différentes expériences présentées dans ce document ont pour la plupart révélé des comportements différents suivant le contenu utilisé. Cette observation est valable pour les tâches de localisation, où la distance perçue est plus ou moins compressée en fonction du stimulus. Les expériences portant sur l'évaluation de la qualité d'expérience ont également révélé une forte dépendance des jugements en fonction de la séquence audio-visuelle présentée. Il serait intéressant d'étudier cette dépendance de manière plus précise.

Enfin, il est important de s'interroger sur la portabilité des résultats de ces travaux à d'autres technologies de restitution visuelle. La télévision 4K, par exemple, semble pouvoir offrir aux spectateurs une sensation de réalisme accrue grâce à l'augmentation de la définition des images. Combinée aux techniques HDR (*High Dynamic Range*) et HFR (*High Frame Rate*), la technologie 4K offre une restitution plus fluide des mouvements et même une certaine sensation de profondeur visuelle. Pour ces raisons, il est tout à fait possible que la restitution de sons spatialisés suivant la distance puisse également contribuer à rendre l'expérience audio-visuelle 4K plus réaliste et immersive.

Bibliographie

[Alais and Burr, 2004] Alais, D. and Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, 14(3) :257–262.

[Alais and Carlile, 2005] Alais, D. and Carlile, S. (2005). Synchronizing to real events : subjective audiovisual alignment scales with perceived auditory depth and speed of sound. In *Proceedings of the National Academy of Sciences of the United States of America*, volume 102, pages 2244–2247.

[Alais et al., 2010] Alais, D., Newell, F. N., and Mamassian, P. (2010). Multisensory processing in review : from physiology to behaviour. *Seeing and Perceiving*, 23(1) :3–38.

[Algazi et al., 2001] Algazi, V. R., Avendano, C., and Duda, R. O. (2001). Elevation localization and head-related transfer function analysis at low frequencies. *J. Acoust. Soc. Am.*, 109(3) :1110–1122.

[Algazi et al., 2002a] Algazi, V. R., Duda, R. O., Duraiswami, R., Gumerov, N. A., and Tang, Z. (2002a). Approximating the head-related transfer function using simple geometric models of the head and torso. *J. Acoust. Soc. Am.*, 112(5) :2053–2064.

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

[Anderson and Zahorik, 2011] Anderson, P. W. and Zahorik, P. (2011). Auditory and visual distance estimation. *Proceedings of Meetings on Acoustics*, 12(1) :050004.

[André et al., 2014] André, C. R., Corteel, E., Embrechts, J.-J., Verly, J. G., and Katz, B. F. (2014). Subjective evaluation of the audiovisual spatial congruence in the case of stereoscopic-3d video and wave field synthesis. *International Journal of Human-Computer Studies*, 72(1) :23–32.

[Arnold et al., 2005] Arnold, D. H., Johnston, A., and Nishida, S. (2005). Timing sight and sound. *Vision research*, 45(10) :1275–1284.

[Ashmead et al., 1995] Ashmead, D. H., Davis, D. L., and Northington, A. (1995). Contribution of listeners' approaching motion to auditory distance perception. *Journal of Experimental Psychology : Human Perception and Performance*, 21(2) :239–256.

[Atal and Schroeder, 1966] Atal, B. S. and Schroeder, M. (1966). Apparent sound source translator. *United States Patent 3,236,949*.

[Balter et al., 2008] Balter, R., Fournier, J., Gicquel, J.-C., Kaptein, R., and Vinayagamoorthy, V. (2008). Technical requirements. *3D4you WP4 - Deliverable 4.1.*

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

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

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

[Berg, 2005] Berg, J. (2005). Evaluation of perceived spatial audio quality. In *Proceedings of the 9th World Multi-Conference on Systemics, Cybernetics and Informatics*, volume 4, pages 10–14.

[Berg and Rumsey, 1999] Berg, J. and Rumsey, F. (1999). Spatial attributes identification and scaling by repertory grid technique and other methods. In *Proceedings of the 16th International Conference on Spatial Sound Reproduction*.

[Berg and Rumsey, 2003] Berg, J. and Rumsey, F. (2003). Systematic evaluation of perceived spatial quality. In *Audio Engineering Society Conference : 24th International Conference : Multichannel Audio, The New Reality*. Audio Engineering Society.

[Berkhout, 1988] Berkhout, A. J. (1988). A holographic approach to acoustic control. *J. Audio Eng. Soc*, 36(12) :977–995.

[Berkhout et al., 1993] Berkhout, A. J., de Vries, D., and Vogel, P. (1993). Acoustic control by wave field synthesis. *The Journal of the Acoustical Society of America*, 93(5) :2764–2778.

[Bertet, 2009] Bertet, S. (2009). *Formats audio 3D hiérarchiques : caractérisation objective et perceptive des systèmes ambisonics d'ordres supérieurs*. PhD thesis, INSA Lyon, Lyon, France.

[Bertet et al., 2006] Bertet, S., Daniel, J., and Moreau, S. (2006). 3d sound field recording with higher order ambisonics-objective measurements and validation of spherical microphone. In *Audio Engineering Society Convention 120*. Audio Engineering Society.

[Blauert, 1970] Blauert, J. (1970). Sound localization in the median plane. *Acustica*, 22 :205–213.

[Blauert, 1997] Blauert, J. (1997). *Spatial hearing : the psychophysics of human sound localization*. MIT press.

[Blumlein, 1933] Blumlein, A. D. (filed Dec. 14, 1931, issued June 14, 1933). Improvements in and relating to sound-transmission, sound-recording and sound-reproducing systems. *British Patent 394,325*.

[Bodenheimer et al., 2007] Bodenheimer, B., Meng, J., Wu, H., Narasimham, G., Rump, B., McNamara, T. P., Carr, T. H., and Rieser, J. J. (2007). Distance estimation in virtual and real environments using bisection. In *Proceedings of the 4th symposium on Applied perception in graphics and visualization*, pages 35–40. ACM.

[Boone et al., 1995] Boone, M. M., Verheijen, E. N. G., and van Tol, P. F. (1995). Spatial sound-field reproduction by wave-field synthesis. *J. Audio Eng. Soc*, 43(12) :1003–1012.

[Bowen et al., 2011] Bowen, A. L., Ramachandran, R., Muday, J. A., and Schirillo, J. A. (2011). Visual signals bias auditory targets in azimuth and depth. *Experimental Brain Research*, 214(3) :403–414.

[Bruneau, 1983] Bruneau, M. (1983). *Introduction aux théories de l'acoustique*. Université du Maine.

[Brungart and Rabinowitz, 1999] Brungart, D. S. and Rabinowitz, W. M. (1999). Auditory localization of nearby sources. head-related transfer functions. *The Journal of the Acoustical Society of America*, 106(3) :1465–1479.

[Brungart and Scott, 2001] Brungart, D. S. and Scott, K. R. (2001). The effects of production and presentation level on the auditory distance perception of speech. *The Journal of the Acoustical Society of America*, 110(1) :425.

[Burr and Alais, 2006] Burr, D. and Alais, D. (2006). Combining visual and auditory information. *Progress in brain research*, 155 :243–258.

[Chen, 2012] Chen, W. (2012). *Caractérisation multidimensionnelle de la qualité d'expérience en télévision de la TV3D stéréoscopique*. PhD thesis, Université de Nantes Angers Le Mans.

[Chen et al., 2010] Chen, W., Fournier, J., Barkowsky, M., Le Callet, P., et al. (2010). New requirements of subjective video quality assessment methodologies for 3dtv. In *Fifth International Workshop on Video Processing and Quality Metrics for Consumer Electronics - VPQM 2010*.

[Choisel and Wickelmaier, 2006] Choisel, S. and Wickelmaier, F. (2006). Extraction of auditory features and elicitation of attributes for the assessment of multichannel reproduced sound. *J. Audio Eng. Soc.*, 54(9) :815–826.

[Coleman, 1962] Coleman, P. D. (1962). Failure to localize the source distance of an unfamiliar sound. *J. Acoust. Soc. Am.*, 34 :345–346.

[Cooper and Bauck, 1989] Cooper, D. H. and Bauck, J. L. (1989). Prospects for transaural recording. *Journal of the Audio Engineering Society*, 37(1/2) :3–19.

[Corrigan et al., 2013] Corrigan, D., Gorzel, M., Squires, J., and Boland, F. (2013). Depth perception of audio sources in stereo 3d environments. In *Proc. SPIE*, volume 8648, pages 864816–1–864816–13.

[Corteel, 2004] Corteel, E. (2004). *Caractérisation et extensions de la wave field synthesis en conditions réelles*. PhD thesis, Université de Paris 6, Paris, France.

[Corteel et al., 2008] Corteel, E., Kuhn-Rahloff, C., and Pellegrini, R. (2008). Wave field synthesis rendering with increased aliasing frequency. In *Audio Engineering Society Convention 124*. Audio Engineering Society.

[Corteel et al., 2012] Corteel, E., Rohr, L., Falourd, X., Nguyen, K.-V., Lissek, H., and others (2012). Practical 3 dimensional sound reproduction using wave field synthesis, theory and perceptual validation. *Acoustics 2012 Nantes*.

[Coutant and Westheimer, 1993] Coutant, B. E. and Westheimer, G. (1993). Population distribution of stereoscopic ability. *Ophthalmic and Physiological Optics*, 13(1) :3–7.

[Craven and Gerzon, 1977] Craven, P. G. and Gerzon, M. (filed July7, 1975, issued Aug. 16, 1977). Coincident microphone simulation covering three dimensionnal space and yielding various directional outputs. *US Patent 4,042,779*.

[Cutting and Vishton, 1995] Cutting, J. and Vishton, P. (1995). Perception of space and motion, chapter perceiving layout and knowing distances : The integration, relative potency, and contextual use of different information about depth. *Academic Pr*, 46 :69–117.

[Côté et al., 2011] Côté, N., Koehl, V., Paquier, M., and Devillers, F. (2011). Interaction between auditory and visual distance cues in virtual reality applications. In *Proceedings of Forum Acusticum 2011*, pages 1275–1280.

[Da Silva, 1985] Da Silva, J. A. (1985). Scales for perceived egocentric distance in a large open field : Comparison of three psychophysical methods. *The American Journal of Psychology*, 98(1) :119.

[Damaske and Wagener, 1969] Damaske, P. and Wagener, B. (1969). Richtungshorversuche uber einen nachgebildeten kopf. *Acustica*, 21 :30–35.

[Daniel, 2000] Daniel, J. (2000). *Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia*. PhD thesis, Ph. D. Thesis, University of Paris VI, France.

[Daniel and Moreau, 2004] Daniel, J. and Moreau, S. (2004). Further study of sound field coding with higher order ambisonics. In *Audio Engineering Society Convention 116*. Audio Engineering Society.

[Daniel et al., 2003] Daniel, J., Moreau, S., and Nicol, R. (2003). Further investigations of high-order ambisonics and wavefield synthesis for holophonic sound imaging. In *Audio Engineering Society Convention 114*. Audio Engineering Society.

[Dixon and Spitz, 1980] Dixon, N. F. and Spitz, L. (1980). The detection of auditory visual desynchrony. *Perception*, 9(6) :719–721.

[Dodgson, 2004] Dodgson, N. A. (2004). Variation and extrema of human interpupillary distance. In *Proceedings of SPIE*, volume 5291, pages 36–46.

[Drhey, 2009] Drhey, A. (2009). *Mormeck, une nouvelle dimension*. Orange Sport Prod.

[Duda and Martens, 1998] Duda, R. O. and Martens, W. L. (1998). Range dependence of the response of a spherical head model. *J. Acoust. Soc. Am.*, 104(5) :3048–3058.

[EBU R-135, 2012] EBU R-135, R. (2012). *EBU-Recommandation R-135 : "Production and exchange formats for 3DTV programmes"*. European Broadcasting Union.

[Ernst and Banks, 2002] Ernst, M. O. and Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870) :429–433.

[Fendrich and Corballis, 2001] Fendrich, R. and Corballis, P. M. (2001). The temporal cross-capture of audition and vision. *Perception & Psychophysics*, 63(4) :719–725.

[Fournier, 1995] Fournier, J. (1995). *Étude de la qualité visuelle des images stéréoscopiques en télévision*. PhD thesis, Université de Rennes I, France.

[Fujisaki et al., 2004] Fujisaki, W., Shimojo, S., Kashino, M., and Nishida, S. (2004). Recalibration of audiovisual simultaneity. *Nature Neuroscience*, 7(7) :773–778.

[Gamespot, 2014] Gamespot (2014). *http ://www.gamespot.com/articles/dead-space-3-us-sales-hit-605000/1100-6405376/.* Page web consultée le 15/12/14.

[Gardner, 1968] Gardner, M. B. (1968). Proximity image effect in sound localization. *The Journal of the Acoustical Society of America*, 43(1) :163.

[Gardner, 1969] Gardner, M. B. (1969). Distance estimation of 0 degree or apparent 0 degree-oriented speech signals in anechoic space. *The Journal of the Acoustical Society of America*, 45(1) :47–53.

[Gardner, 1997] Gardner, W. G. (1997). *3-D audio using loudspeakers.* PhD thesis, Kluwer Academic Publishers.

[Gerzon, 1973] Gerzon, M. A. (1973). Periphony : With-height sound reproduction. *Journal of the Audio Engineering Society*, 21(1) :2–10.

[Gerzon, 1985] Gerzon, M. A. (1985). Ambisonics in multichannel broadcasting and video. *J. Audio Eng. Soc*, 33(11) :859–871.

[Godfroy et al., 2003] Godfroy, M., Roumes, C., and Dauchy, P. (2003). Spatial variations of visual-auditory fusion areas. *Perception*, 32(10) :1233–1246.

[Gogel and Tietz, 1973] Gogel, W. C. and Tietz, J. D. (1973). Absolute motion parallax and the specific distance tendency. *Perception & Psychophysics*, 13(2) :284–292.

[Gorzel et al., 2012] Gorzel, M., Corrigan, D., Kearney, G., Squires, J., and Boland, F. (2012). Distance perception in virtual audio-visual environments. In *25th AES UK Conference : Spatial Audio in Today's 3D World.* Audio Engineering Society.

[Guillon, 2009] 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.* PhD thesis, Université du Maine, Le Mans, France.

[Hafter and De Maio, 1975] Hafter, E. and De Maio, J. (1975). Difference thresholds for interaural delay. *The Journal of the Acoustical Society of America*, 57(1) :181–187.

[Hamasaki et al., 2004] Hamasaki, K., Hatano, W., and Hiyama, K. (2004). 5.1 and 22.2 multichannel sound productions using an integrated surround sound panning system. *AES 117th Convention.*

[Hartnagel, 2007] Hartnagel, D. (2007). *La perception de l'espace multisensoriel appréhendée par l'étude de la fusion visuo-auditive : les effets de la dissociation des référentiels spatiaux.* PhD thesis, Paris 8.

[Hebrank and Wright, 1974] Hebrank, J. and Wright, D. (1974). Spectral cues used in the localization of sound sources on the median plane. *J. Acoust. Soc. Am.*, 56(6) :1829–1834.

[Heron et al., 2007] Heron, J., Whitaker, D., McGraw, P. V., and Horoshenkov, K. V. (2007). Adaptation minimizes distance-related audiovisual delays. *Journal of Vision*, 7(13) :5.

[Hershkowitz and Durlach, 1969] Hershkowitz, R. M. and Durlach, N. I. (1969). Interaural time and amplitude jnds for a 500-hz tone. *The Journal of the Acoustical Society of America*, 46(6B) :1464–1467.

[Hládek et al., 2013] Hládek, L., Le Dantec, C. C., Kopčo, N., and Seitz, A. (2013). Ventriloquism effect and aftereffect in the distance dimension. In *Proceedings of Meetings on Acoustics*, volume 19, page 050042. Acoustical Society of America.

[Hofman et al., 1998] Hofman, P. M., Van Riswick, J. G., and Van Opstal, A. J. (1998). Relearning sound localization with new ears. *Nature neuroscience*, 1(5) :417–421.

[Hollier and Rimell, 1998] Hollier, M. P. and Rimell, A. N. (1998). An experimental investigation into multi-modal synchronization sensitivity for perceptual model development. In *Audio Engineering Society Convention 105*.

[Howard, 1982] Howard, I. P. (1982). *Human visual orientation*. Wiley, Chichester.

[Howard and Templeton, 1966] Howard, I. P. and Templeton, W. B. (1966). Human spatial orientation.

[Iosono, 2014] Iosono (2014). *http ://www.iosono-sound.com/game-audio/*. Page web consultée le 15/12/14.

[ISO/IEC DIS 23008-3, 2014] ISO/IEC DIS 23008-3, . (2014). *Draft International Standard MPEG-H : High efficiency coding and media delivery in heterogeneous environments — Part 3 : 3D audio*. ISO/IEC Working Group 11.

[ITU 1116, 2014] ITU 1116, R. (2014). *ITU-Recommandation BS 1116-2 : "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems"*. International Telecommunications Union, Radio-communication Assembly.

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

[ITU 1286, 1997] ITU 1286, R. (1997). *ITU-Recommandation BS 1286 : "Methods for the subjective assessment of audio systems with accompanying picture"*. International Telecommunications Union, Radio-communication Assembly.

[ITU 1534, 2014] ITU 1534, R. (2014). *ITU-Recommandation BS 1534-2 : "Method for the subjective assessment of intermediate quality level of coding systems"*. International Telecommunications Union, Radio-communication Assembly.

[ITU 1788, 2007] ITU 1788, R. (2007). *ITU-Recommandation BT 1788 : "Methodology for the subjective assessment of video quality in multimedia applications"*. International Telecommunications Union, Radio-communication Assembly.

[ITU 2020, 2014] ITU 2020, R. (2014). *ITU-Recommandation BT 2020-1 : "Parameter values for ultra-high definition television systems for production and international programme exchange"*. International Telecommunications Union, Radio-communication Assembly.

[ITU 2021, 2012] ITU 2021, R. (2012). *ITU-Recommandation BT 2021 : "Interactive test methods for audiovisual communications"*. International Telecommunications Union, Radio-communication Assembly.

[ITU 2022, 2012] ITU 2022, R. (2012). *ITU-Recommandation BT 2022 : "General viewing conditions for subjective assessment of quality of SDTV and HDTV television pictures on flat panel displays"*. International Telecommunications Union, Radio-communication Assembly.

[ITU 500, 2012] ITU 500, R. (2012). *ITU-Recommandation BT 500-13 : "Methodology for the subjective assessment of the quality of television pictures"*. International Telecommunications Union, Radio-communication Assembly.

[ITU 775, 2012] ITU 775, R. (2012). *ITU-Recommandation BS 775-3 : "Multichannel stereophonic sound system with and without accompagnying picture"*. International Telecommunications Union, Radio-communication Assembly.

[ITU E800, 2008] ITU E800, R. (2008). *ITU-Recommandation T E800 : "Definitions of terms related to quality of service"*. International Telecommunications Union, Radio-communication Assembly.

[ITU P10, 2008] ITU P10, R. (2008). *ITU-Recommandation T P10 : "Vocabulary for performance and quality of service : Amendment 2"*. International Telecommunications Union, Radio-communication Assembly.

[ITU P910, 2008] ITU P910, R. (2008). *ITU-Recommandation T P910 : "Subjective video quality assessment methods for multimedia applications"*. International Telecommunications Union, Radio-communication Assembly.

[ITU P911, 1998] ITU P911, R. (1998). *ITU-Recommandation T P911 : "Subjective audiovisual quality assessment methods for multimedia applications"*. International Telecommunications Union, Radio-communication Assembly.

[Jones and McManus, 1986] Jones, B. L. and McManus, P. R. (1986). Graphic scaling of qualitative terms. *SMPTE journal*, 95(11) :1166–1171.

[Jot et al., 1999] Jot, J.-M., Larcher, V., and Pernaux, J.-M. (1999). A comparative study of 3-d audio encoding and rendering techniques. In *Audio Engineering Society Conference : 16th International Conference : Spatial Sound Reproduction*. Audio Engineering Society.

[Kearney et al., 2012] Kearney, G., Gorzel, M., Rice, H., and Boland, F. (2012). Distance perception in interactive virtual acoustic environments using first and higher order ambisonic sound fields. *Acta Acustica united with Acustica*, 98(1) :61–71.

[Kim and Choi, 2005] Kim, S.-M. and Choi, W. (2005). On the externalization of virtual sound images in headphone reproduction : A wiener filter approach. *J. Acoust. Soc. Am.*, 117(6) :3657–3665.

[Kirkeby et al., 1998a] Kirkeby, O., Nelson, P. A., and Hamada, H. (1998a). Local sound field reproduction using two closely spaced loudspeakers. *The Journal of the Acoustical Society of America*, 104(4) :1973–1981.

[Kirkeby et al., 1998b] Kirkeby, O., Nelson, P. A., and Hamada, H. (1998b). The "stereo dipole" - a virtual source imaging system using two closely spaced loudspeakers. *J. Audio Eng. Soc.*, 46(5) :387–395.

[Kistler and Wightman, 1992] Kistler, D. J. and 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(3) :1637–1647.

[Klein et al., 2009] Klein, E., Swan, J. E., Schmidt, G. S., Livingston, M. A., and Staadt, O. G. (2009). Measurement protocols for medium-field distance perception in large-screen immersive displays. In *Virtual Reality Conference, 2009. VR 2009. IEEE*, pages 107–113.

[Klumpp and Eady, 1956] Klumpp, R. and Eady, H. (1956). Some measurements of interaural time difference thresholds. *The Journal of the Acoustical Society of America*, 28(5) :859–860.

[Kohlrausch and van de Par, 2005] Kohlrausch, A. and van de Par, S. (2005). Audio-visual interaction in the context of multi-media applications. In Blauert, J., editor, *Communication Acoustics*, pages 109–138. Springer Berlin Heidelberg.

[Komiyama, 1989] Komiyama, S. (1989). Subjective evaluation of angular displacement between picture and sound directions for hdtv sound systems. *Journal of the Audio Engineering Society*, 37(4) :210–214.

[Komiyama et al., 1991] Komiyama, S., Morita, A., Kurozumi, K., and Nakabayashi, K. (1991). Distance control system for a sound image. In *Audio Engineering Society Conference : 9th International Conference : Television Sound Today and Tomorrow*. Audio Engineering Society.

[Kopčo et al., 2004] Kopčo, N., Schoolmaster, M., and Shinn-Cunningham, B. (2004). Learning to judge distance of nearby sounds in reverberant and anechoic environments. In *Proc. Joint congress CFA/DAGA*.

[Lappin et al., 2006] Lappin, J. S., Shelton, A. L., and Rieser, J. J. (2006). Environmental context influences visually perceived distance. *Perception & psychophysics*, 68(4) :571–581.

[Larcher, 2001] Larcher, V. (2001). *Techniques de spatialisation des sons pour la réalité virtuelle*. PhD thesis, Université de Paris VI, France.

[Le Bagousse et al., 2010] Le Bagousse, S., Paquier, M., and Colomes, C. (2010). Families of sound attributes for assessment of spatial audio. In *Audio Engineering Society Convention 129*.

[Letowski, 1989] Letowski, T. (1989). Sound quality assessment : concepts and criteria. *AES 87th Convention*.

[Lewald, 2002] Lewald, J. (2002). Rapid adaptation to auditory-visual spatial disparity. *Learning & Memory*, 9(5) :268–278.

[Lewald et al., 2001] Lewald, J., Ehrenstein, W. H., and Guski, R. (2001). Spatio-temporal constraints for auditory–visual integration. *Behavioural Brain Research*, 121(1) :69–79.

[Loomis et al., 1998] Loomis, J. M., Klatzky, R. L., Philbeck, J. W., and Golledge, R. G. (1998). Assessing auditory distance perception using perceptually directed action. *Perception & Psychophysics*, 60(6) :966–980.

[Loomis and Knapp, 2003] Loomis, J. M. and Knapp, J. M. (2003). Visual perception of egocentric distance in real and virtual environments. *Virtual and adaptive environments*, (11) :21–46.

[Marins et al., 2008] Marins, P., Rumsey, F., and Zielinski, S. (2008). Unravelling the relationship between basic audio quality and fidelity attributes in low bit-rate multi-channel audio codecs. In *Audio Engineering Society Convention 124*. Audio Engineering Society.

[Marschall and Chang, 2013] Marschall, M. and Chang, J. (2013). Sound-field reconstruction performance of a mixed-order ambisonics microphone array. *Proceedings of Meetings on Acoustics*, 19 :1–9.

[McAnally and Martin, 2002] McAnally, K. I. and Martin, R. L. (2002). Variability in the headphone-to-ear-canal transfer function. *J. Audio Eng. Soc.*, 50(4) :263–266.

[McGurk and MacDonald, 1976] McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264 :746–748.

[Mershon et al., 1980] Mershon, D. H., Desaulniers, D. H., Amerson, T. L., and Kiefer, S. A. (1980). Visual capture in auditory distance perception : Proximity image effect reconsidered. *Journal of Auditory Research*, 20(2) :129–136.

[Mershon and King, 1975] Mershon, D. H. and King, L. E. (1975). Intensity and reverberation as factors in the auditory perception of egocentric distance. *Perception & Psychophysics*, 18(6) :409–415.

[Middlebrooks, 1992] Middlebrooks, J. (1992). Narrow-band sound localization related external ears acoustics. *J. Acoust. Soc. Am.*, 92(5) :2607–2624.

[Middlebrooks, 1999] Middlebrooks, J. C. (1999). Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency. *The Journal of the Acoustical Society of America*, 106(3) :1493–1510.

[Mills, 1958] Mills, A. W. (1958). On the minimum audible angle. *The Journal of the Acoustical Society of America*, 30(4) :237–246.

[Mills, 1960] Mills, A. W. (1960). Lateralization of high-frequency tones. *The Journal of the Acoustical Society of America*, 32(1) :132–134.

[Moller, 1992] Moller, H. (1992). Fundamentals of binaural technology. *Applied Acoustics*, 36(3-4) :171–218.

[Moore, 2003] Moore, B. (2003). An introduction to the psychology of hearing.

[Moreau, 2006] Moreau, S. (2006). *Etude et réalisation d'outils avancés d'encodage spatial pour la technique de spatialisation sonore Higher Order Ambisonics : microphone 3D et contrôle de la distance*. PhD thesis, Ph. D. Thesis, Université du Mans, Le Mans, France.

[Mullin et al., 2001] Mullin, J., Smallwood, L., Watson, A., and Wilson, G. (2001). New techniques for assessing audio and video quality in real-time interactive communication. In *Proceedings of IHM-HCI*.

[Munoz Soto et al., 2008] Munoz Soto, R., Recuero, M., Duran, D., and Gazzo, M. (2008). Absolute threshold of coherence of position perception between auditory and visual sources for dialogs. In *Audio Engineering Society Convention 125*. Audio Engineering Society.

[Nicol, 2010] Nicol, R. (2010). *Représentation et perception des espaces auditifs virtuels*. PhD thesis, Mémoire d'Habilitation à Diriger des Recherches, Université du Mans, Le Mans, France.

[Nicol et al., 2013] Nicol, R., Emerit, M., Gros, L., Pallone, G., Palacino, J., and Moulin, S. (2013). Le son 3d dans les futurs services de télécommunication. *Acoustique & Techniques*, 71 :4–9.

[Noisternig et al., 2013] Noisternig, M., Carpentier, T., and Warusfel, O. (2013). Dispositif de spatialisation sonore 3d à l'espace de projection de l'ircam - un réseau de 345 haut-parleurs pour une restitution par wfs et hoa. *Acoustique & Techniques*, 71 :30–39.

[Palmer, 1999] Palmer, S. (1999). *Vision Science : Photons to Phenomenology*. Bradford book.

[Patterson, 2007] Patterson, R. (2007). Human factors of 3-d displays. *Journal of the Society for Information Display*, 15(11) :861–871.

[Pernaux et al., 1998] Pernaux, J.-M., Boussard, P., and Jot, J.-M. (1998). Virtual sound source positioning and mixing in 5.1 implementation on the real-time system genesis. In *Proc. Conf. Digital Audio Effects (DAFx-98)*, page 76–80. Citeseer.

[Perrott, 1993] Perrott, D. R. (1993). Auditory and visual localization : Two modalities, one world. In *Audio Engineering Society 12th International Conference : The Perception of Reproduced Sound*.

[Pulkki, 1997] Pulkki, V. (1997). Virtual sound source positioning using vector base amplitude panning. *Journal of the Audio Engineering Society*, 45(6) :456–466.

[Rayleigh, 1907] Rayleigh, L. (1907). On our perception of sound direction. *Philisophical Magazine*, 13 :214–232.

[Renard, 2000] Renard, C. (2000). *Analyse objective et subjective d'une technique de rendu sonore 2D sur une zone d'écoute étendue, l'holophonie, en vue de réaliser un mur de télé-présence*. PhD thesis, Master Thesis, Université du Mans, Le Mans, France.

[Renner et al., 2013] Renner, R. S., Velichkovsky, B. M., and Helmert, J. R. (2013). The perception of egocentric distances in virtual environments - a review. *ACM Computing Surveys*, 46(2) :1–40.

[Rohr et al., 2013] Rohr, L., Corteel, E., Nguyen, K.-V., and Lissek, H. (2013). Vertical localization performance in a practical 3-d wfs formulation. *Journal of the Audio Engineering Society*, 61(12) :1001–1014.

[Rosenblum et al., 1987] Rosenblum, L. D., Carello, C., and Pastore, R. E. (1987). Relative effectiveness of three stimulus variables for locating a moving sound source. *Perception*, 16(2) :175–186.

[Ruch and Fulton, 1960] Ruch, T. C. and Fulton, J. F. (1960). *Medical physiology and biophysics*, volume 35. LWW.

[Rugeles et al., 2014] Rugeles, F., Emerit, M., and Katz, B. F. (2014). Evaluation objective et subjective de différentes méthodes de lissage des hrtf. In *Congres Français d'Acoustique - CFA 2014*, pages 2219–2221.

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

[Rumsey et al., 2005] Rumsey, F., Zielinski, S., Kassier, R., and Bech, S. (2005). On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality. *The Journal of the Acoustical Society of America*, 118(2) :968–976.

[Rébillat, 2011] Rébillat, M. (2011). *Vibrations de plaques multi-excitateurs de grandes dimensions pour la création d'environnements virtuels audio-visuels*. PhD thesis, Ecole polytechnique, France.

[Rébillat et al., 2012] Rébillat, M., Boutillon, X., Corteel, E., and Katz, B. F. G. (2012). Audio, visual, and audio-visual egocentric distance perception by moving subjects in virtual environments. *ACM Transactions on Applied Perception*, 9(4) :1–17.

[Schroeder and Atal, 1963] Schroeder, M. and Atal, B. (1963). Computer simulation of sound transmission in rooms. *Proceedings of the IEEE*, 51(3) :536–537.

[Sekuler et al., 1997] Sekuler, R., Sekuler, A., and Lau, R. (1997). Sound alters visual motion perception. *Nature*, 385(6614) :308–308.

[Shams et al., 2000] Shams, L., Kamitani, Y., and Shimojo, S. (2000). Illusions : What you see is what you hear. *Nature*, 408(6814) :788–788.

[Shaw et al., 1991] Shaw, B. K., McGowan, R. S., and Turvey, M. (1991). An acoustic variable specifying time-to-contact. *Ecological Psychology*, 3(3) :253–261.

[Shaw and Teranishi, 1968] Shaw, E. and 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(1) :240–249.

[Shinn-Cunningham, 2000a] Shinn-Cunningham, B. G. (2000a). Distance cues for virtual auditory space. In *Proceedings of the IEEE-PCM*, volume 2000, pages 227–230.

[Shinn-Cunningham, 2000b] Shinn-Cunningham, B. G. (2000b). Learning reverberation : Considerations for spatial auditory displays. In *Proceedings of the International Conference on Auditory Displays*, pages 126–134.

[Shinn-Cunningham et al., 2005] Shinn-Cunningham, B. G., Kopco, N., and Martin, T. J. (2005). Localizing nearby sound sources in a classroom : Binaural room impulse responses. *The Journal of the Acoustical Society of America*, 117(5) :3100–3115.

[Shipley, 1964] Shipley, T. (1964). Auditory flutter-driving of visual flicker. *Science*, 145(3638) :1328–1330.

[Spector, 1990] Spector, R. H. (1990). *Visual Fields- In : Clinical Methods : The History, Physical, and Laboratory Examinations*. Walker HK, Hall WD, Hurst JW, editors.

[Spence et al., 2003] Spence, C., Baddeley, R., Zampini, M., James, R., and Shore, D. I. (2003). Multisensory temporal order judgments : when two locations are better than one. *Perception & psychophysics*, 65(2) :318–328.

[Spors et al., 2009] Spors, S., Wierstorf, H., Geier, M., and Ahrens, J. (2009). Physical and perceptual properties of focused sources in wave field synthesis. In *127th AES Convention*.

[Start, 1997] Start, E. (1997). *Direct sound enhancement by Wave Field Synthesis*. PhD thesis, Ph. D. Thesis, Delft University of Technology, Delft, The Netherlands.

[Stevens, 1957] Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, 64(3) :153–181.

[Stone et al., 2001] Stone, J., Hunkin, N., Porrill, J., Wood, R., Keeler, V., Beanland, M., Port, M., and Porter, N. (2001). When is now ? perception of simultaneity. *Proceedings of the Royal Society of London. Series B : Biological Sciences*, 268(1462) :31–38.

[Sugita and Suzuki, 2003] Sugita, Y. and Suzuki, Y. (2003). Audiovisual perception : Implicit estimation of sound-arrival time. *Nature*, 421(6926) :911–911.

[Susini et al., 1999] Susini, P., McAdams, S., and Winsberg, S. (1999). A multidimensional technique for sound quality assessment. *Acustica*, 85 :1105–1115.

[Teghtsoonian and Teghtsoonian, 1969] Teghtsoonian, M. and Teghtsoonian, R. (1969). Scaling apparent distance in natural indoor settings. *Psychonomic Science*, 16(6) :281–283.

[Thurlow and Jack, 1973] Thurlow, W. R. and Jack, C. E. (1973). Certain determinants of the "ventriloquism effect". *Perceptual and motor skills*, 36(3c) :1171–1184.

[Ukai and Howarth, 2008] Ukai, K. and Howarth, P. A. (2008). Visual fatigue caused by viewing stereoscopic motion images : Background, theories, and observations. 29(2) :106–116.

[Vatakis and Spence, 2007] Vatakis, A. and Spence, C. (2007). Crossmodal binding : Evaluating the "unity assumption" using audiovisual speech stimuli. *Perception & Psychophysics*, 69(5) :744–756.

[Vogel, 1993] Vogel, P. (1993). *Application of Wave Field Synthesis in room acoustics*. PhD thesis, Ph. D. Thesis, Delft University of Technology, Delft, The Netherlands.

[Walker and Scott, 1981] Walker, J. T. and Scott, K. J. (1981). Auditory-visual conflicts in the perceived duration of lights, tones, and gaps. *Journal of Experimental Psychology : Human Perception and Performance*, 7(6) :1327–1339.

[Welch, 1999] Welch, R. B. (1999). Chapter 15 meaning, attention, and the "unity assumption" in the intersensory bias of spatial and temporal perceptions. In Gisa Aschersleben, T. B. and Müsseler, J., editors, *Cognitive Contributions to the Perception of Spatial and Temporal Events*, volume 129 of *Advances in Psychology*, pages 371–387. North-Holland.

[Welch and Warren, 1980] Welch, R. B. and Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological bulletin*, 88(3) :638–667.

[Wettschurek, 1971] Wettschurek, R. (1971). Uber unterschiedsschwellen beim richtungshören in der medianebene. *Fortschritte der Akustik – DAGA 70*, pages 385–388.

[Wierstorf et al., 2013] Wierstorf, H., Raake, A., Geier, M., and Spors, S. (2013). Perception of focused sources in wave field synthesis. *Journal of the Audio Engineering Society*, 61(1/2) :5–16.

[Wierstorf et al., 2012] Wierstorf, H., Raake, A., and Spors, S. (2012). Localization of a virtual point source within the listening area for wave field synthesis. In *Audio Engineering Society Convention 133*. Audio Engineering Society.

[Wiest and Bell, 1985] Wiest, W. M. and Bell, B. (1985). Stevens's exponent for psychophysical scaling of perceived, remembered, and inferred distance. *Psychological Bulletin*, 98(3) :457.

[Wittek et al., 2004] Wittek, H., Kerber, S., Rumsey, F., and Theile, G. (2004). Spatial perception in wave field synthesis rendered sound fields : Distance of real and virtual nearby sources. In *116th AES Convention*.

[Woods, 2001] Woods, A. J. (2001). Optimal usage of LCD projectors for polarized stereoscopic projection. In *Photonics West 2001-Electronic Imaging*, pages 5–7. International Society for Optics and Photonics.

[Woodworth and Schloesberg, 1962] Woodworth, R. S. and Schloesberg, G. (1962). *Experimental psychology*. New-York.

[Yamanoue et al., 2006] Yamanoue, H., Okui, M., and Okano, F. (2006). Geometrical analysis of puppet-theater and cardboard effects in stereoscopic HDTV images. *IEEE Transactions on Circuits and Systems for Video Technology*, 16(6) :744–752.

[Yamanoue et al., 2000] Yamanoue, H., Okui, M., and Yuyama, I. (2000). A study on the relationship between shooting conditions and cardboard effect of stereoscopic images. *Circuits and Systems for Video Technology, IEEE Transactions on*, 10(3) :411–416.

[Yano et al., 2004] Yano, S., Emoto, M., and Mitsuhashi, T. (2004). Two factors in visual fatigue caused by stereoscopic HDTV images. *Displays*, 25(4) :141–150.

[Yarrow et al., 2011] Yarrow, K., Jahn, N., Durant, S., and Arnold, D. H. (2011). Shifts of criteria or neural timing ? the assumptions underlying timing perception studies. *Consciousness and cognition*, 20(4) :1518–1531.

[Yost and Dye Jr, 1988] Yost, W. A. and Dye Jr, R. H. (1988). Discrimination of interaural differences of level as a function of frequency. *The Journal of the Acoustical Society of America*, 83(5) :1846–1851.

[Zahorik, 2001] Zahorik, P. (2001). Estimating sound source distance with and without vision. *Optometry and Vision Science*, 78(5) :270–275.

[Zahorik, 2002a] Zahorik, P. (2002a). Assessing auditory distance perception using virtual acoustics. *The Journal of the Acoustical Society of America*, 111(4) :1832–1846.

[Zahorik, 2002b] Zahorik, P. (2002b). Auditory display of sound source distance. In *Proceedings of the 8th International Conference on Auditory Displays*, page 326–332.

[Zahorik et al., 2005] Zahorik, P., Brungart, D. S., and Bronkhorst, A. W. (2005). Auditory distance perception in humans : A summary of past and present research. *Acta Acustica united with Acustica*, 91(3) :409–420.

[Zielinski et al., 2007] Zielinski, S., Brooks, P., and Rumsey, F. (2007). On the use of graphic scales in modern listening tests. *AES 123rd Convention*.

[Zwislocki and Feldman, 1956] Zwislocki, J. and Feldman, R. S. (1956). Just noticeable differences in dichotic phase. *The Journal of the Acoustical Society of America*, 28(5) :860–864.