Influence de la stéréoscopie sur la perception du son - Cas des mixages sonores pour le cinéma en relief

Etienne Hendrickx

université de bretagne occidentale

THÈSE / UNIVERSITÉ DE BRETAGNE OCCIDENTALE
sous le sceau de l'Université européenne de Bretagne

pour obtenir le titre de
DOCTEUR DE L'UNIVERSITÉ DE BRETAGNE OCCIDENTALE

Mention : STIC
Spécialité : Acoustique

École Doctorale SICMA

présentée par

Etienne Hendrickx

Préparée au Laboratoire en Sciences et Techniques de l'Information, de la Communication et de la Connaissance (Lab-STICC), Brest

Influence de la stéréoscopie sur la perception du son – Cas des mixages sonores pour le cinéma en relief

Thèse soutenue le 4 décembre 2015
devant le jury composé de :

Gilles COPPIN
Professeur, Télécom Bretagne / *directeur de thèse*

Brian F.G. KATZ
Directeur de recherche, LIMSI - CNRS / *rapporteur*

Rozenn NICOL
Ingénieur, Orange Labs / *rapporteur*

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

Etienne PARIZET
Professeur, INSA Lyon / *examinateur*

Isabelle VIAUD-DELMON
Directeur de recherche, IRCAM / *examinateur*

Vincent KOEHL
Maître de conférences, Université de Bretagne Occidentale / *invité*

CROSS CHANNEL FILM LAB

france (manche channel) england interreg

European Regional Development Fund
The European Union, investing in your future

Fonds européen de développement régional
L'union Européenne investit dans votre avenir

Remerciements

Tout d'abord, je tiens à remercier Rozenn Nicol et Brian F.G. Katz, qui ont accepté d'être rapporteurs de cette thèse. Je remercie également Etienne Parizet et Isabelle Viaud-Delmon d'avoir accepté de participer au jury.

Un grand merci également à Gilles Coppin, qui a accepté de diriger mes travaux de recherche. Merci, bien évidemment, à mon encadrant Mathieu Paquier, pour sa bonne humeur et son soutien permanent. Nos débats sur la dimension métaphysique de l'ANOVA, sur l'enseignement des musiques actuelles dans les conservatoires, et sur les différences entre une clarinette et un hautbois vont bien me manquer...

Je tiens évidemment à remercier sincèrement toute l'équipe de l'Université de Brest, à commencer par Vincent Koehl, pour ses relectures attentives d'articles, ses sempiternelles boutades sur Tatayet et l'effet ventriloque et surtout pour ses incroyables tableaux à double entrée. Merci également à Erwan Le Morvan, qui a rendu possible les deux expériences au cinéma « Le Bretagne » de Saint-Renan, à Vincent Mazo, Julian Palacino, François Barvec, Luc Pennamen et bien évidemment à Marie, Patricia et Isabelle.

Un très grand merci à Pierre Souchar, qui a accepté de filmer une grande partie des séquences utilisées dans cette thèse. Je remercie également toute l'équipe du CCFL (Cross Channel Film Lab), Antoine Le Bos, Estelle Car et Benoît Méline, ainsi que les réalisateurs Pablo Agüero, Henry Davies, Gaëlle Denis, Joséphine Derobe et mon compatriote concarnois Giil Taws, d'avoir accepté que j'utilise des extraits de leurs films pour mes tests.

Je remercie tous les sujets qui ont participé aux tests subjectifs, en particulier ceux qui ont dû affronter l'expérience IV, dans laquelle Baptiste Le Deun (que je remercie par la même occasion) répétait inlassablement la même phrase pendant 4 heures... Quel courage !

Bien évidemment, un grand merci à ma famille pour leur confiance et leurs encouragements. Et pour finir, je remercie du fond du cœur Claire pour son soutien sans faille, ses livraisons de chez *Histoire de Chocolat* quand le moral vacillait, ses relectures attentives, et sa bonne humeur permanente et inaltérable, même quand je rentrais grognon le soir (probablement exaspéré par un bug insolvable dans LaTeX).

Cette recherche s'inscrit dans le projet « Cross Channel Film Lab 2 », sélectionné dans le cadre du programme européen de coopération transfrontalière INTERREG IV A France (Manche) - Angleterre, cofinancé par le FEDER.

Certaines études de cette thèse ont également été soutenues par l'Agence Nationale de la Recherche dans le cadre du projet EDISON 3D (ANR-13-CORD-0008-02).

Résumé

Peu d'études ont été menées sur l'influence de la stéréoscopie sur la perception d'un mixage audio au cinéma. Les témoignages de mixeurs ou les articles scientifiques montrent pourtant une grande diversité d'opinions à ce sujet. Certains estiment que cette influence est négligeable, d'autres affirment qu'il faut totalement revoir notre conception de la bande-son, aussi bien au niveau du mixage que de la diffusion.

Une première série d'expériences s'est intéressée à la perception des sons d'ambiance. 8 séquences, dans leurs versions stéréoscopiques (3D-s) et non-stéréoscopiques (2D), ont été diffusées dans un cinéma à des sujets avec plusieurs mixages différents. Pour chaque présentation, les sujets devaient évaluer à quel point le mixage proposé leur paraissait trop frontal ou au contraire trop « surround », le but étant de mettre en évidence une éventuelle influence de la stéréoscopie sur la perception de la balance frontal/surround d'un mixage audio. Les résultats obtenus ont rejoint ceux d'une expérience préliminaire menée dans un auditorium de mixage, où les sujets se trouvaient en situation de mixeur et devaient eux-mêmes régler la balance frontal/surround : l'influence de la stéréoscopie était faible et n'apparaissait que pour quelques séquences. Une troisième expérience fut conduite pour vérifier si les séquences pour lesquelles la perception de la balance frontal/surround était significativement impactée par la stéréoscopie étaient celles dont les différences entre versions 2D et 3D-s étaient les plus importantes en termes de profondeur visuelle perçue. Cependant, aucune corrélation n'a pu être trouvée.

Des études ont ensuite été menées sur la perception des objets sonores tels que dialogues et effets. Une quatrième expérience s'est intéressée à l'effet ventriloque en élévation : lorsque l'on présente à un sujet des stimuli audio et visuel temporellement coïncidents mais spatialement disparates, les sujets perçoivent parfois le stimulus sonore au même endroit que le stimulus visuel. On appelle ce phénomène l'*effet ventriloque* car il rappelle l'illusion créée par le ventriloque lorsque sa voix semble plutôt provenir de sa marionnette que de sa propre bouche. Ce phénomène a été très largement étudié dans le plan horizontal, et dans une moindre mesure en distance. Par contre, très peu d'études se sont intéressées à l'élévation. Dans cette expérience, nous avons présenté à des sujets des séquences audiovisuelles montrant un homme en train de parler. Sa voix pouvait être reproduite sur différents haut-parleurs, qui créaient des disparités plus ou moins grandes en azimut et en élévation entre le son et l'image. Pour chaque présentation, les sujets devaient indiquer si la voix semblait ou non provenir de la même direction que la bouche de l'acteur. Les résultats ont montré que l'effet ventriloque était très efficace en élévation, ce qui suggère qu'il n'est peut-être pas nécessaire de rechercher la cohérence

audiovisuelle en élévation au cinéma.

Une cinquième et dernière expérience a permis d'étudier l'influence de la stéréoscopie sur les attentes des spectateurs en termes de cohérence audiovisuelle spatiale. Au cinéma, les objets sonores sont en général diffusés sur l'enceinte centrale, indépendamment de la position à l'écran des sources visuelles associées. Cependant, certains ingénieurs du son et chercheurs ont suggéré que la cohérence audiovisuelle spatiale pouvait améliorer significativement l'expérience des spectateurs, surtout dans le cas de films en relief. Dans cette expérience, les sujets devaient évaluer à quel point la bande-son leur paraissait « adaptée » à l'image pour 8 séquences projetées en 2D et en 3D-s. Selon la bande-son, les sources sonores pouvaient être plus ou moins cohérentes en azimut et en profondeur avec la position de leur source visuelle respective sur l'écran (la cohérence en élévation avait été mise de côté au vu des résultats de l'expérience 4). Les résultats ont montré que la cohérence en azimut pouvait améliorer significativement l'adéquation du son à l'image. En profondeur, une amélioration a pu être constatée, mais seulement pour une séquence. Par contre, la stéréoscopie n'a eu aucune influence sur les jugements des sujets, en accord avec les résultats des premières expériences sur la perception des sons d'ambiance.

Abstract

The influence of stereoscopy on sound perception - a case study on the sound mixing of stereoscopic-3D movies

Few psychoacoustic studies have been carried out about the influence of stereoscopy on the sound mixing of movies. Yet very different opinions can be found in the cinema industry and in scientific papers. Some argue that sound needs to be mixed differently for stereoscopic movies while others pretend that this influence is negligible.

A first set of experiments was conducted, which focused on the perception of ambience. Eight sequences - in their stereoscopic (s-3D) and non-stereoscopic (2D) versions, with several different sound mixes - were presented to subjects. For each presentation, subjects had to judge to what extent the mix sounded frontal or « surround. » The goal was to verify whether stereoscopy had an influence on the perception of the front/surround balance of ambience. Results showed that this influence was weak, which was consistent with a preliminary experiment conducted in a mixing auditorium where subjects had to mix the front/surround balance of several sequences themselves.

A third experiment was conducted to verify if the sequences that were significantly influenced by stereoscopy corresponded to the sequences whose differences between s-3D and 2D versions were the most important in terms of perceived visual depth, yet no substantial correlation could be found.

Studies were then conducted on the perception of sound objects such as dialogs or on-screen effects. A fourth experiment focused on ventriloquism in elevation : when presented with a spatially discordant auditory-visual stimulus, subjects sometimes perceive the sound and the visual stimuli as coming from the same location. Such a phenomenon is often referred to as *ventriloquism*, because it evokes the illusion created by a ventriloquist when his voice seems to emanate from his puppet rather than from his mouth. While this effect has been extensively examined in the horizontal plane and to a lesser extent in distance, few psychoacoustic studies have focused on elevation. In this experiment, sequences of a man talking were presented to subjects. His voice could be reproduced on different loudspeakers, which created disparities in both azimuth and elevation between the sound and the visual stimuli. For each presentation, subjects had to indicate whether or not the voice seemed to emanate from the mouth of the actor. Ventriloquism was found to be highly effective in elevation, which suggests that audiovisual coherence in elevation might be unnecessary in theaters.

In a fifth experiment, the influence of stereoscopy on subjects' expectations regarding

audiovisual spatial coherence was investigated. In theaters, sound objects are most of the time reproduced on the central loudspeaker, regardless of the position on screen of their related visual sources. Yet, some sound engineers and researchers have suggested that a spatial audiovisual coherence could improve the experience of the audience significantly, especially for s-3D movies. In this experiment, subjects were asked to judge the suitability of several soundtracks for 8 sequences, which were presented in their s-3D and 2D versions. Depending on the soundtrack, sound sources could be more or less coherent in azimuth and in depth to their related visual sources (coherence in elevation was not investigated because of the results of the fourth experiment). Results showed that sound suitability could be significantly improved for most of the sequences when coherence in azimuth was achieved. In depth, improvement was only observed with one sequence. However, no significant effect of stereoscopy on subjects' judgments could be found, which is consistent with the previous experiments on the perception of ambience.

Table des matières

Introduction

L'image 3D impacte significativement l'expérience audiovisuelle des spectateurs par rapport à une image 2D. Il est donc naturel de se demander si cet impact ne modifie pas également la perception ou les attentes des spectateurs concernant la bande-son. Pour l'instant, les pratiques de mixage et les technologies de reproduction sonore qui accompagnent les films 3D-stéréoscopiques (3D-s) au cinéma sont restées proches voire identiques à celles utilisées pour les films 2D. Il est cependant possible que d'autres pratiques ou d'autres technologies permettent une diffusion du son en plus grande adéquation avec les spécificités de l'image en relief.

Peu d'études ont jusqu'à maintenant exploré cette problématique. Certaines ont comparé différents systèmes de reproduction sonore pour tenter de déterminer le système le plus adapté à l'image 3D-s. Cependant, nous pensons qu'il est important avant tout de **vérifier si l'influence de la stéréoscopie sur la perception du son existe bel et bien**. Les témoignages d'ingénieurs du son ayant mixé des films en 3D-s sont plutôt contradictoires : certains estiment que cette influence est négligeable, d'autres affirment qu'il faut revoir notre conception de la bande-son, aussi bien au niveau du mixage que de la diffusion.

Si l'influence de la stéréoscopie s'avère significative, il faudra tenter d'en appréhender la nature et d'en évaluer l'importance. Ainsi, en comprenant mieux comment la stéréoscopie change notre perception et nos attentes sonores **par rapport à une projection « classique » en 2D**, peut-être sera-t-il plus aisé de proposer des pistes d'amélioration, ou de prédire si certaines propositions récentes (systèmes de spatialisation WFS, *Dolby Atmos*, *Auro-3D*, etc.) semblent adéquates.

Nous regarderons donc dans des cas de figures très simples si l'image provoque chez le spectateur des perceptions ou des attentes sonores différentes, selon qu'elle est projetée en 2D ou en 3D-stéréoscopique. L'ambition du projet CCFL est d'explorer les possibilités de la 3D-s pour le cinéma d'auteur, comme l'a fait Wim Wenders dans son film *Pina*. Nous écarterons donc de notre champ d'étude les films à gros budget (tels que *Avatar* de James Cameron), qui intègrent souvent scènes d'action, effets de jaillissement et images de synthèse, et nous concentrerons sur des exemples "moins spectaculaires", issus de productions à petit ou moyen budget. Il est aussi à noter que nous nous limiterons dans le cadre de cette thèse à l'étude de configurations associées au cinéma « main stream », et non à d'autres applications (télévision,

réalité virtuelle, jeux vidéo, installations expérimentales, évènementiel, etc.) également susceptibles d'associer une présentation visuelle stéréoscopique à des systèmes de spatialisation sonore.

Pour Alfonso Cuarón, réalisateur de *Gravity*, la bande-son du film participe autant que la 3D à l'immersion du spectateur. Le son a donc un rôle important à jouer pour convaincre les spectateurs d'aller découvrir les films en relief au cinéma plutôt que sur leur ordinateur ou leur tablette numérique, ce qui contribuerait à enrayer le déclin du cinéma 3D auquel nous assistons depuis quelques années, et ce malgré le succès récent de *Gravity*.

Organisation du manuscrit

Le manuscrit présenté ici s'articule autour de deux parties principales :

— La Première partie propose un État de l'Art des connaissances scientifiques et techniques nécessaires pour atteindre les objectifs de la thèse :

- Le premier chapitre se focalise sur la perception et la diffusion du son. Dans un premier temps, nous présentons les mécanismes et performances de la localisation auditive. Nous proposons ensuite un historique des principaux systèmes de reproduction sonore utilisés au cinéma, de la naissance du cinéma parlant jusqu'à aujourd'hui. Nous en profiterons pour faire un point sur les pratiques de mixage actuelles. Nous présentons également des systèmes de spatialisation du son plus récents, susceptibles de s'imposer au cinéma dans un avenir proche et de remettre en question certaines de ces pratiques ;
- Le deuxième chapitre se focalise sur la perception visuelle et sur l'image stéréoscopique. Dans un premier temps, nous présentons les mécanismes et performances de la localisation visuelle. Nous décrivons ensuite les principes de la stéréoscopie, de la captation à la diffusion. Enfin, nous terminons par un bref historique du cinéma en relief ;
- Le troisième chapitre aborde différents phénomènes d'interactions entre le son et et l'image susceptibles d'influencer l'expérience des spectateurs au cinéma. Dans un premier temps, nous présentons l'*effet ventriloque*. Dans un deuxième temps, nous proposons un état de l'art des études ayant été conduites plus spécifiquement sur la perception du son lié à l'image en 3D-s.

À l'issue de cet État de l'Art, nous préciserons nos hypothèses.

— La deuxième partie présente les contributions de la thèse, qui s'articulent autour de 5 expériences :

- Le quatrième chapitre présente les expériences I, II et III, dans lesquelles nous avons essayé de mettre en évidence l'influence de la stéréoscopie sur la perception des sons d'ambiance (ambiances de mer, de ville, etc.) ;

- Le cinquième chapitre présente les expériences IV et V, dans lesquelles nous avons souhaité mettre en évidence l'influence de la stéréoscopie sur les attentes des spectateurs en termes de spatialisation des objets sonores (i.e. dialogues et effets sonores ponctuels tels que bruits de pas, porte qui claque, etc.).

Première partie

État de l'Art

Chapitre 1

Le son

Sommaire

Ce chapitre se focalise sur la perception et la diffusion du son. Dans un premier temps, nous présentons les mécanismes et performances de la localisation auditive. Nous proposons ensuite un historique des principaux systèmes de reproduction sonore utilisés au cinéma, de la naissance du cinéma parlant jusqu'à aujourd'hui. Nous en profiterons pour faire un point sur les pratiques de mixage actuelles. Nous présentons également des systèmes de spatialisation du son plus récents, susceptibles de s'imposer au cinéma dans un avenir proche et de remettre en question certaines de ces pratiques.

1.1 La localisation auditive

1.1.1 En azimut

Lorsqu'un son vient de droite, il arrive d'abord à notre oreille droite puis ensuite à notre oreille gauche. Il en résulte une différence de temps, communément appelée ITD (*interaural time difference*), entre les signaux des deux oreilles. En effet, si on modélise la tête humaine par une simple sphère de rayon a, alors un son provenant de droite avec un azimut θ doit

parcourir une distance supplémentaire à peu près égale à $a \sin\theta + a\theta$ pour arriver à l'oreille gauche (voir Fig. 1.1) , ce qui correspond à un retard de $a(\sin\theta + \theta)/c$ si c désigne la célérité du son dans l'air (Woodworth et Schlosberg, 1962).

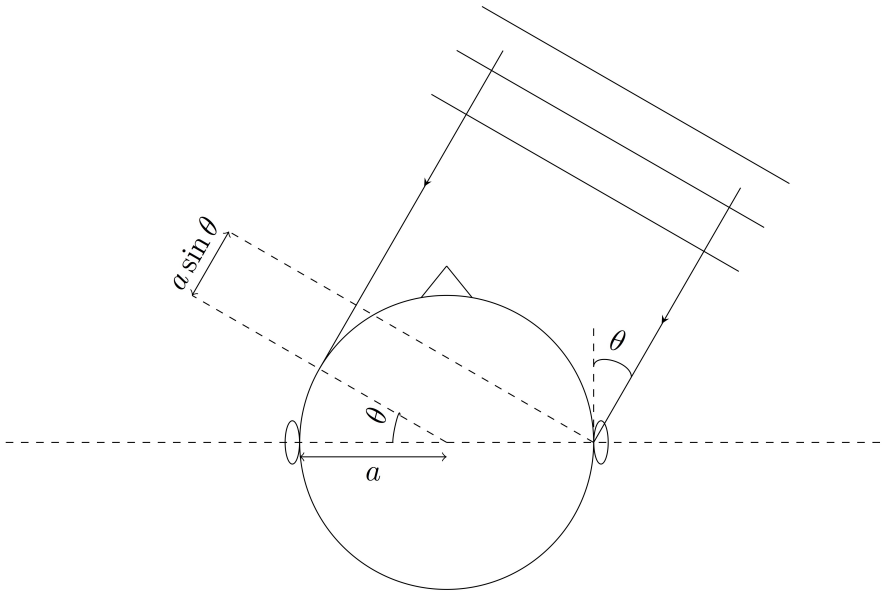


FIGURE 1.1 – Onde sonore plane atteignant une sphère de rayon a modélisant la tête. Le signal doit parcourir une distance plus longue pour atteindre l'oreille gauche, d'où un retard à peu près égal à $a(\sin\theta + \theta)/c$. D'après Stern *et al.* (2006).

Selon Blauert (1997) et Moore (2012), ces différences de temps caractéristiques peuvent donner des informations sur la position de la source sonore jusqu'à une fréquence d'environ 1500 Hz. Au-delà, la longueur d'onde est trop petite devant la distance entre les deux oreilles ($\approx$ 17 cm) et l'azimut θ ne peut plus être déduit de la différence de phase entre les deux oreilles, car cette dernière est supérieure à 2π.

A des fréquences plus élevées, la tête de l'auditeur, grande devant la longueur d'onde, constitue un obstacle et occasionne des phénomènes de diffractions et de réflexions plus ou moins importants selon la direction et la fréquence de l'onde incidente. Ainsi, un son venant de droite donne lieu à une pression sonore plus faible au niveau de l'oreille gauche que de l'oreille droite (Stewart, 1920). Il en résulte une différence de niveau, communément appelée ILD (*interaural level difference*), qui constitue un indice important de la localisation à des fréquences supérieures à 1500 Hz.

1.1.2 En élévation

L'ITD et l'ILD, malgré leur importance primordiale, ne renseignent que sur l'azimut d'une source et ne donnent pas d'information sur l'élévation. En effet, à des valeurs d'ITD et d'ILD égales peuvent correspondre plusieurs élévations possibles dans l'espace. Les indices spectraux (IS) permettent dans une certaine mesure de lever ces incertitudes : en arrivant sur le corps de l'auditeur, les ondes sonores subissent des réflexions et des diffractions sur les pavillons des

oreilles, le torse, les épaules, etc. Il en résulte une coloration spécifique du son, dépendante de l'angle d'incidence du son. Selon Middlebrooks (1992) et Guillon (2009), la comparaison de cette coloration spécifique avec d'autres filtrages stockés en mémoire permet d'estimer l'élévation d'une source.

L'ILD, l'ITD et les indices spectraux sont caractérisés par une fonction appelée Head Related Transfer Function (HRTF). Ces filtres HRTF sont définis pour chaque oreille et chaque position de la source sonore (azimut, élévation et distance). Cependant, pour des distances supérieures à 1 m, les filtres HRTF ne sont que peu influencés par la distance (Duda et Martens, 1998).

1.1.3 En distance

La localisation en distance utilise des mécanismes différents de ceux utilisés pour la perception de la direction. Mershon and King [1975] ont défini quatre indices acoustiques de perception de la distance : l'intensité, la réverbération, le contenu spectral et les différences binaurales. Les deux premiers indices sont ceux qui varient le plus avec la distance et les plus importants d'un point de vue perceptif (Shinn-Cunningham, 2000).

Intensité

L'intensité est souvent considérée comme l'indice principal de perception de la distance d'une source sonore, car la variation de niveau sonore est aisément détectée par un auditeur. L'intensité donne des informations sur la distance car elle décroit au fur et à mesure que la distance à la source augmente. Cette décroissance est fonction des propriétés de l'environnement et de la source : par exemple, pour 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 (Bruneau, 1983).

Le niveau sonore est un indice relatif : il ne permet pas d'estimer la distance absolue à laquelle se situe la source, à moins de connaître a priori le niveau de la source sonore. Des études ont en effet montré que l'estimation de la distance égocentrique d'une source sonore en champ libre était difficile (Gardner, 1969), mais qu'elle pouvait être considérablement améliorée si les sujets étaient familiarisés avec la source (Mershon et King, 1975).

Réverbération

La perception de la distance entre la source sonore et l'auditeur dépend également du ratio entre l'énergie du champ direct et celle du champ réverbéré. La réverbération est déterminée par la forme et les propriétés acoustiques des murs de la salle d'écoute et par les objets situés dans la salle d'écoute. 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.

Il a été montré qu'une augmentation de la réverbération du lieu d'écoute augmentait la distance perçue de la source sonore (Butler *et al.*, 1980; Nielsen, 1993; Mershon et Bowers, 1979). Dans une expérience rapportée par Warren (1999), une personne parlait distinctement devant un microphone tout en s'éloignant progressivement. L'amplification avait ensuite été réglée de façon à ce que l'intensité des sons délivrés par les haut-parleurs aux sujets soit perçue comme constante. Malgré l'absence d'indice d'intensité, l'impression d'éloignement était parfaitement conservée car la réverbération se faisait de plus en plus forte.

Il s'agit en général d'un indice relatif, moins précis que l'intensité pour estimer la distance relative, ou pour détecter que la position d'une source sonore a changé (Bronkhorst et Houtgast, 1999). Cependant, la familiarisation avec la réverbération d'une salle peut permettre à l'indice de devenir absolu (Mershon et King, 1975).

Contenu spectral

Le contenu spectral intervient de deux façons dans la perception de la distance :

— Pour un objet situé à une distance supérieure à 15 m, l'absorption de l'onde acoustique dans l'air va provoquer une atténuation, notamment dans les hautes fréquences (Coleman, 1968). A l'inverse, un signal avec beaucoup de hautes fréquences sera perçu comme plus proche de l'auditeur.

— En fonction des propriétés d'absorption des murs d'une salle, qui varient en fonction de la fréquence, la réverbération va plus ou moins renforcer ou atténuer certaines zones spectrales, et donc colorer le son d'autant plus que le champ réverbéré sera prédominant par rapport au champ direct.

Comme pour le niveau sonore, la coloration spectrale n'apporte qu'une information restreinte sur la position de la source sonore si le spectre du son n'est pas connu a priori. Il s'agit donc d'un indice de localisation relatif.

Indices binauraux

Lorsque la source sonore est proche de l'auditeur, l'auditeur est capable d'extraire une information de distance à partir des deux indices ITD et ILD (Coleman, 1968). Ces différences binaurales sont particulièrement efficaces dans un environnement anéchoïque, où elles permettent une évaluation de la distance de la source de manière absolue. Cependant, elles n'apportent pas beaucoup d'informations supplémentaires pour des conditions réverbérées ou pour des distances égocentriques supérieures à 1 m (Shinn-Cunningham *et al.*, 2005; Brungart et Rabinowitz, 1999).

1.1.4 Performances de localisation

Les performances de localisation du système auditif peuvent être mesurées de différentes manières :

— dans une tâche de localisation absolue, les sujets doivent indiquer la position dans l'espace d'une source sonore :
  - l'écart entre la position moyenne perçue et la position réelle de la source (ou *erreur de localisation*) renseigne sur l'*acuité* du système auditif ;
  - la dispersion des résultats de part et d'autre de la position moyenne renseigne sur la *précision* du système auditif. L'écart-type est souvent utilisé pour mesurer la précision (Letowski et Letowski, 2011).

— dans une tâche de localisation relative, deux sons successifs sont en général présentés aux sujets. Ces derniers doivent alors rapporter s'ils ont perçu les deux sons à la même position. En augmentant progressivement la distance entre les deux sons à partir d'une même position initiale, une courbe psychométrique peut être mesurée :
  - l'angle pour lequel 50% des participants perçoivent un changement de position est alors appelé MAA (« minimum audible angle », aussi appelée *Flou de localisation* par Blauert (1997)) ;
  - la largeur de la fonction psychométrique peut également être mesurée et est parfois appelée *acuité* (Alais et Burr, 2004), ce qui peut provoquer des confusions par rapport à la définition que nous avons donnée de ce terme dans le cas d'une tâche de localisation absolue.

Recanzone a montré que le flou de localisation obtenu à partir d'une tâche de localisation relative ne permet pas de prédire l'erreur de localisation qui serait obtenue avec une tâche de localisation absolue. Par contre, la largeur d'une fonction psychométrique issue d'une tâche de localisation relative est de l'ordre de la largeur d'une distribution obtenue à partir d'une tâche de localisation absolue (Recanzone *et al.*, 1998).

En azimut

Preibisch-Effenberger (1966) et Haustein et Schirmer (1970), dans deux expériences synthétisées par Blauert (1997), ont utilisé des salves de bruit blanc pour étudier la localisation absolue dans le plan horizontal (voir Fig. 1.2). L'erreur de localisation était la plus petite devant et derrière le sujet (environ 1°) et bien plus grande sur les côtés (environ 10°). Des résultats semblables ont été obtenus par Carlile *et al.* (1997).

Le flou de localisation en azimut suit la même tendance que l'erreur de localisation, même s'il est légèrement plus grand derrière (5,5°) que devant (3,6°). Les performances se dégradent progressivement lorsque les sources se rapprochent de l'axe interaural, avec un flou atteignant 10° sur les côtés. D'autres études ont montré que le flou de localisation dans la zone frontale

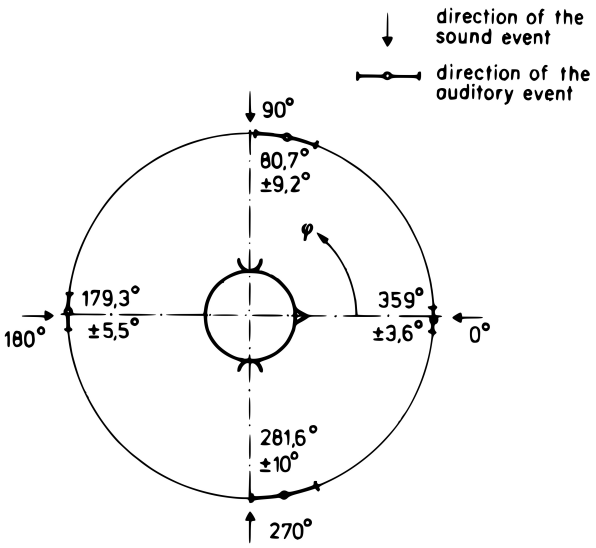


FIGURE 1.2 – Erreurs et flous de localisation dans le plan horizontal avec des salves de bruit blanc (Preibisch-Effenberger, 1966; Haustein et Schirmer, 1970). D'après Blauert (1997).

pouvait varier de 1° à 5° selon la nature des stimuli. Par exemple, Perrott et Saberi (1990) ont obtenu un flou de localisation de 0,97° avec des successions de clics.

Les performances de localisation dépendent également de la fréquence, et les maxima d'erreurs et de flous de localisation se retrouvent souvent à des fréquences comprises entre 1.5 kHz et 3 kHz. Ce phénomène pourrait être expliqué par la faiblesse des indices de localisation dans cette zone fréquentielle, puisque les fréquences y sont trop élevées pour que l'ITD soit efficace et en même temps trop basses pour que l'ILD soit efficace (Middlebrooks et Green, 1991).

Selon Oldfield et Parker (1984), l'erreur de localisation en azimut serait indépendante de l'élévation.

En élévation

En élévation, les performances de localisation devraient être moindres puisque le système auditif ne dispose plus des indices ITD et ILD et doit se reposer sur les indices spectraux, moins fiables, pour évaluer la hauteur des sources.

Les études montrent en effet des erreurs et flous de localisation plus grands en élévation qu'en azimut. Carlile *et al.* (1997), par exemple, ont obtenu des erreurs moyennes de 4° avec des salves de bruit pleine-bande. Oldfield et Parker (1984) ont obtenu à partir de bruits blancs des erreurs de localisation pouvant atteindre 8 degrés en élévation contre 6 degrés en azimut pour des sources frontales. Quant à Perrott et Saberi (1990), ils ont obtenu à partir de séries de clics des flous de localisation égaux à 3,65° en élévation contre 0,97° en azimut.

Les flous de localisation varient grandement en fonction de la source : 17° avec la voix d'une personne inconnue (Blauert, 1970), 9° avec une voix familière (Damaske et Wagener, 1969)

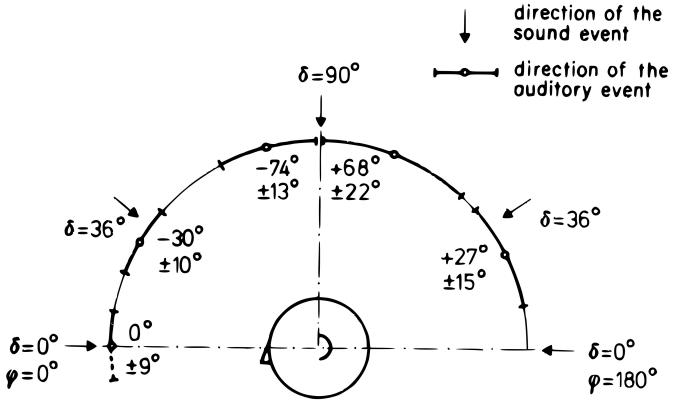


FIGURE 1.3 – Erreurs et flous de localisation dans le plan médian avec une voix familière (Damaske et Wagener, 1969). D'après Blauert (1997).

(voir Fig. 1.3), et 4° avec un bruit blanc (Wettschurek, 1970). Ces résultats suggèrent qu'une certaine familiarisation avec la source est nécessaire pour obtenir de bonnes performances de localisation en élévation.

Damaske et Wagener (1969) ont obtenu des erreurs et des flous de localisation qui augmentaient avec l'élévation (voir Fig. 1.3), contrairement à Oldfield et Parker (1984) dont les erreurs étaient indépendantes de l'élévation.

La précision de localisation est également moins bonne en élévation qu'en azimut : lorsque Makous et Middlebrooks (1990) ont demandé à des sujets de localiser des bruits pleine-bande dont la position pouvait varier en azimut et en élévation, la variabilité intra-sujet des réponses (c'est-à-dire l'écart-type des réponses de part et d'autre de la réponse moyenne pour un seul et même individu) était en moyenne 2.5 fois plus importante en élévation qu'en azimut.

En distance

Zahorik (2002a) a observé que la distance des sources proches (<1,5m) était systématiquement surestimée tandis que celle des sources lointaines étaient sous-estimée.

A partir des résultats de 84 études et de la loi de Stevens (1957) (loi de puissance permettant de relier l'intensité d'un stimulus à la sensation qu'il produit sur le sujet), Zahorik a obtenu une bonne approximation de la relation entre distance perçue et distance physique des sources sonores :

$$D' = 1,32 \times D^{0,54}$$

où D' est la distance perçue et D la distance réelle (Zahorik *et al.*, 2005). La fonction est tracée sur la Figure 1.4 et montre bien que les distances proches sont effectivement surestimées tandis que les distances lointaines sont sous-estimées.

Zahorik a également observé une variabilité des jugements de distance importante puis-

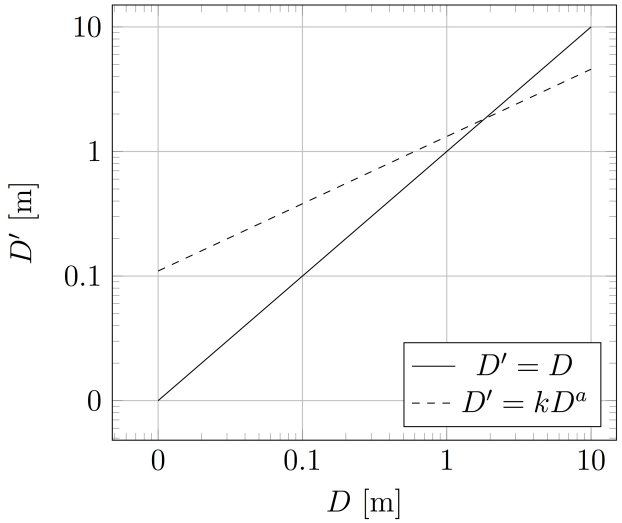


FIGURE 1.4 – Distance perçue D' en fonction de la distance réelle D. Echelle logarithmique. La ligne continue correspond au cas idéal où $D' = D$ tandis que la ligne en pointillées correspond à l'approximation proposée par Zahorik *et al.* (2005) : $D' = 1,32 \times D^{0,54}$. D'après André (2013).

qu'elle atteint 20 à 60% de la distance réelle de la cible suivant les participants (Zahorik, 2002b).

D'autres études suggèrent que les performances peuvent être considérablement améliorées si les sujets sont familiers, ou familiarisés, avec les stimuli. Gardner (1969) a étudié pour des distances allant de 0.9 à 9 mètres la localisation d'une voix d'homme chuchotant, parlant normalement et criant. La voix parlée normalement était correctement localisée, par contre les distances étaient respectivement sous-estimée et sur-estimée pour les voix chuchotée et criée. De bonnes performances ont également été rapportées par Haustein (1969) avec des sons brefs (voir Fig. 1.5), qui avaient auparavant été répétitivement présentés aux sujets à différentes distances.

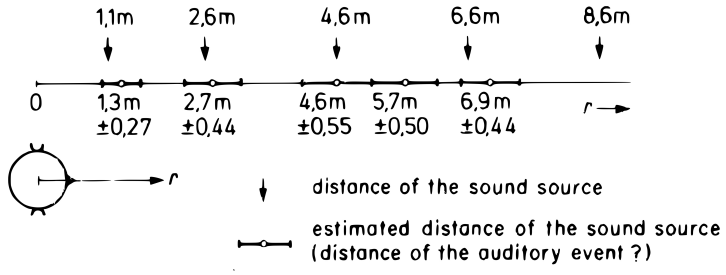


FIGURE 1.5 – Erreurs et flous de localisation dans la distance avec des sons brefs (Haustein, 1969). D'après Blauert (1997).

1.2 Reproduction sonore spatialisée

Traditionnellement, on sépare les sons d'un film en quatre catégories :

— les dialogues ;

— les musiques, qui peuvent être soit :

- diégétiques, c'est à dire qu'elles émanent d'une source située directement dans le lieu et le temps de l'action, par exemple une radio posée sur une table, ou un acteur se mettant à jouer d'un instrument ;
- extra-diégétiques, c'est-à-dire qu'elles accompagnent l'image depuis une position *off*, en dehors du lieu et du temps de l'action. On parle souvent de « Musique de fosse » pour la musique extra-diégétique, en référence à la fosse d'orchestre de l'opéra classique.

— les effets : bruits de pas, coup de feu, bruit d'un verre posé sur une table par l'acteur, etc. ;

— les ambiances sonores : pluie, vent, rumeur de ville, etc.

Dans ce chapitre, nous allons remonter le cours du temps et voir comment le traitement de ces différentes catégories de son a évolué au gré des avancées technologiques :

— dans un premier temps, nous exposerons le principe de la stéréophonie, mis au point par Blumlein (1933) ;

— dans un deuxième temps, nous présenterons les différents systèmes de diffusion qui se sont succédés depuis les débuts du cinéma parlant jusqu'à aujourd'hui, et qui n'ont finalement été jusqu'à maintenant que des extensions de la stéréophonie ;

— nous en profiterons alors pour faire un point sur les pratiques de mixage actuelles ;

— enfin, nous présenterons de nouveaux systèmes de spatialisation, soit en développement (WFS), soit déjà implémentés dans certains cinémas (*Dolby Atmos, Auro-3D*).

1.2.1 La stéréophonie

Dans une configuration stéréophonique classique, deux haut-parleurs forment avec l'auditeur un triangle équilatéral (voir Fig. 1.6). Dans ce cas :

— si les signaux diffusés sur les deux enceintes ne présentent aucune corrélation, alors le sujet perçoit deux sources distinctes sur chacune des deux enceintes, une à $\theta = -30°$ et l'autre à $\theta = +30°$;

— si un même signal est diffusé sur les deux enceintes (synchrone et au même niveau), alors le sujet perçoit une source « fantôme » au centre ($\theta = 0°$), entre les deux haut-parleurs ;

— si un même signal est diffusé sur les deux enceintes, mais avec une différence de niveau, alors la source « fantôme » résultante est perçue plus proche de l'enceinte la plus forte ;

— si un même signal est diffusé sur les deux enceintes, mais avec une différence de temps,

alors la source « fantôme » résultante est perçue plus proche de l'enceinte qui est en avance ;

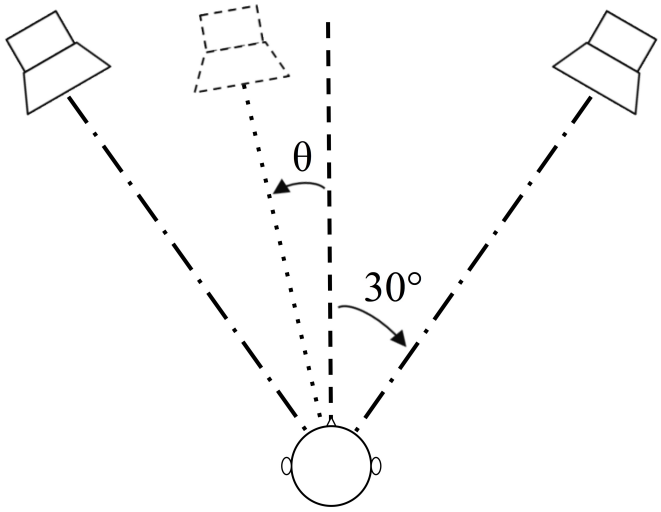


FIGURE 1.6 – Configuration stéréophonique classique, avec une source « fantôme » reproduite à un azimut θ.

Différences de temps et d'intensité peuvent se combiner. Par exemple, pour une source « fantôme » pas trop excentrée ($-23° < \theta < +23°$), l'azimut S peut être estimé à partir de la relation :

$$S(\Delta t, \Delta L) = 44 \times \Delta t + 2.2 \times \Delta L$$

avec S, Δt et ΔL respectivement exprimés en degrés, en ms et en dB (Theile, 2000).

En jouant sur les différences de temps et/ou d'intensité, on peut donc étaler plusieurs sources fantômes dans la latéralité et restituer une scène sonore virtuelle, néanmoins délimités par les deux haut-parleurs. En effet, au-delà de 17 dB d'écart ou de 1,1 ms de décalage entre les deux enceintes, une source est perçue « plaquée » contre l'une des deux enceintes.

En pratique, il existe deux façons d'exploiter la technique stéréophonique :

— naturelle : une scène sonore peut être enregistrée avec deux microphones. Si les microphones sont directifs, alors les signaux enregistrés contiennent naturellement des différences d'intensité. Si les microphones sont non coïncidents, alors les signaux enregistrés contiennent naturellement des différences de temps (Hugonnet et Walder, 1995).

— artificielle : un même signal monophonique peut être diffusé sur les deux haut-parleurs mais avec des gains différents . L'azimut de la source fantôme résultante est alors définie par

$$\frac{\tan\theta}{\tan 30^\circ} = \frac{g_L - g_R}{g_L + g_R}$$

g_R et g_L désignant les gains appliqués respectivement aux haut-parleurs droit et gauche (Pulkki et Karjalainen, 2001). On pourrait également envisager une latéralisation des sources monophoniques par différence de temps, cependant la quasi-totalité des tables de mixages repose sur des différences de niveau (Griesinger, 2002).

Au cinéma, on rencontre à la fois de la stéréophonie naturelle et artificielle :

— les sons d'ambiance et les musiques sont souvent composés d'enregistrements réalisés avec des couples de microphones ;

— les dialogues et les effets sonores monophoniques, lorsqu'ils sont spatialisés, utilisent des différences de niveau entre enceintes (panning d'amplitude). On parle d'« objets sonores » pour désigner de tels sons.

Dans le cas d'une stéréophonie avec panning d'amplitude, la source « fantôme » a son propre ILD, mais l'ITD et les indices spectraux restent ceux des 2 haut-parleurs. Le rendu stéréophonique est 1-D car les sources sonores ne peuvent être rendues que sur une ligne entre les deux haut-parleurs. En effet :

— les sources sont reproduites dans le plan horizontal : il n'y a donc pas d'élévation ;

— le rendu de la distance est fixé aux positions des enceintes. Les indices monauraux de perception de la distance (intensité, coloration spectrale, réverbération) peuvent cependant être simulés pour donner une impression de profondeur.

1.2.2 La diffusion du son au cinéma : un bref historique

Les premiers systèmes multicanaux : d'Abel Gance au *Dolby Stéréo*

Les premières expériences de diffusion multicanale sont apparues très rapidement après l'arrivée du son au cinéma (on considère *The Jazz Singer* d'Alan Crosland (1927) comme étant le premier « film parlant », c'est-à-dire avec dialogues synchrones à l'image). Dès 1932, le réalisateur Abel Gance déposa un brevet pour un système de « projection sonore à haut-parleurs multiples ». Il utilisa ce système dans son film *Napoléon* en 1934, en diffusant lors des scènes de batailles des sons de chevaux, d'explosions et de cris dans des haut-parleurs répartis tout autour de la salle.

6 ans plus tard, en 1940, Walt-Disney présenta *Fantasia* avec le système *Fantasound*. La bande-son se composait de trois canaux alimentant trois enceintes situées derrière l'écran : enceinte gauche, enceinte centrale, enceinte droite. Les canaux gauche et droite alimentaient également des enceintes « surround » situées tout autour de la salle.

Après la seconde guerre mondiale, le système *Cinérama* apparut avec 6 canaux (5 canaux pour les enceintes derrière l'écran et 1 canal supplémentaire pour des enceintes placées tout autour de la salle). La Fox proposa ensuite le *CinémaScope* sous deux versions différentes :

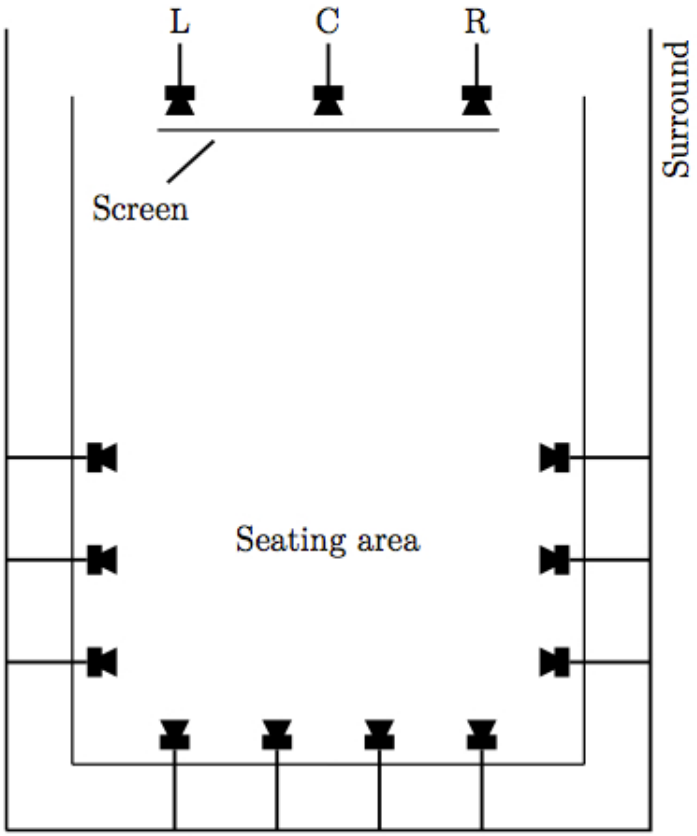


FIGURE 1.7 – Arrangement traditionnel des enceintes dans une salle de cinéma (Streicher et Everest, 2006). Certains formats prévoient l'ajout de 2 enceintes frontales (entre les enceintes gauche et centrale, et les enceintes centrale et droite). Les « surround » peuvent être divisés en 1, 2, 3 ou 4 canaux (voire plus) selon les formats.

4 canaux (pistes magnétiques sur pellicules 35 mm) et 6 canaux (pistes magnétiques sur pellicules 70 mm). Dans les deux cas, un canal était dédié aux enceintes « surround » et les canaux restants alimentaient les enceintes derrière l'écran, comme pour le *Cinérama*. Des tests furent alors réalisés pour que la voix des acteurs suive leurs déplacements à l'image. Malheureusement, le simple champ/contre-champ provoquait une alternance gauche/droite incessante et perturbante pour le spectateur. Il fut alors convenu de toujours diffuser les dialogues sur l'enceinte centrale, les ambiances et la musique étant diffusées dans les autres enceintes (Gambier, 2010).

Cependant, le couchage sur bandes magnétiques était une opération onéreuse, et il fallut attendre *Star Wars*, en 1976, pour que le son surround puisse véritablement se généraliser, avec l'utilisation d'un nouvel encodage permettant d'enregistrer 4 canaux différents sur les deux pistes optiques d'une pellicule standard 35mm : le *Dolby Stéréo*. Ces canaux se divisaient en trois canaux frontaux (gauche, centre, droite) et un canal « surround » (voir Fig. 1.7). On remarque que la ligne d'enceintes surround commence au 1/3 de la salle dans sa longueur et qu'il y a donc un « trou » entre l'écran et les premières enceintes surround (trou que le format *Dolby Atmos* se proposera de combler par la suite).

L'ère du son numérique

Presque vingt ans plus tard, *Batman Returns* de Tim Burton (1992), puis *Jurassic Park* de Steven Spielberg (1993) ouvrèrent l'ère du son numérique au cinéma avec les formats *Dolby Digital* et *DTS* (*Digital Theater System*). Ces deux formats utilisaient la disposition 5.1 :

— 3 canaux alimentaient trois enceintes frontales (gauche, centre, droite) ;

— les « surround » étaient désormais séparés en 2 canaux distincts (« surround gauche » et « surround droit ») ;

— le « .1 » désignait un sixième canal alimentant un ou plusieurs caissons de basse pour les fréquences inférieures à 120 Hz.

Dolby explique avoir privilégié le plan horizontal dans sa configuration 5.1, car les sources sonores sont plus susceptibles d'apparaître dans ce plan-là qu'au-dessus ou en-dessous des spectateurs (Allen, 1991).

Par la suite, d'autres formats apparurent, tels que le *SDDS* 7.1 (Sony Dynamic Digital Sound) proposé par Sony et inauguré en 1993 avec le film *Last Action Hero* de John McTiernan. Deux enceintes frontales supplémentaires (1 enceinte gauche-centre et 1 enceinte centre-droite) devaient permettre de ne plus percevoir de « trous » entre les enceintes gauche, centrale, et droite et de fluidifier les mouvements de sources sonores dans la latéralité. Le système fut cependant un échec commercial et la production fut arrêtée dans les années 2000.

En 1999, Dolby et THX proposèrent également le *Dolby Digital Surround EX* (6.1) avec le film *Star Wars Episode I : The Phantom Menace* de George Lucas. Ce nouveau format proposait un canal surround central en plus du 5.1 traditionnel.

Enfin, le *Dolby Surround 7.1*, avec *Toy Story 3* de Lee Unkrich (2010), proposa de diviser les « surround » en 4 canaux (voir Fig. 1.8) :

— surround latéral gauche ;

— surround latéral droit ;

— surround arrière gauche ;

— surround arrière droit.

Cependant, le 5.1 reste encore aujourd'hui le format le plus répandu à travers le monde.

Tous les formats multicanaux présentés ci-dessus partagent les mêmes caractéristiques que la stéréophonie à 2 canaux : dans le cas d'un panning d'amplitude, la source « fantôme » a son propre ILD, mais l'ITD et les indices spectraux restent ceux des haut-parleurs. Il s'agit également de systèmes 1-D, car les sources sonores ne peuvent se déplacer qu'en azimut. Il n'y a pas d'élévation et le rendu de la distance est fixé aux positions des enceintes.

1.2.3 Pratiques de mixage actuelles

La généralisation du 5.1 pour le mixage des films a entraîné une certaine uniformisation des pratiques de mixage et de spatialisation du son.

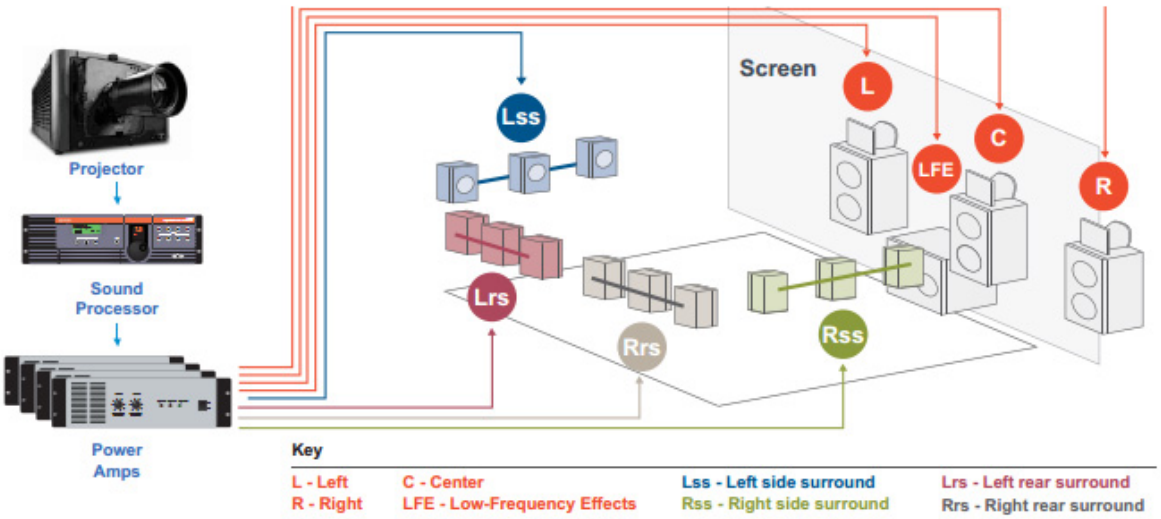


FIGURE 1.8 – Configuration Dolby Surround 7.1 (Dolby, 2011).

Dialogues et effets sonores

En azimut

La plupart du temps, les ingénieurs du son ne latéralisent ni les dialogues ni les effets sonores et les envoient directement dans l'enceinte centrale (Toole, 2008). Une disparité peut donc apparaître en azimut entre une source sonore (par exemple, la voix de l'acteur) et sa source visuelle associée (la bouche de l'acteur) si la position à l'écran de la source visuelle ne coïncide pas avec l'enceinte centrale. Cette pratique de « mixer au centre » est également valable lorsque l'acteur est hors-champ (même s'il existe quelques exceptions où les dialogues sont alors diffusés dans les surround : *Les Autres* d'Alejandro Amenábar, *Gravity* d'Alfonso Cuarón, etc.). Les professionnels avancent plusieurs arguments pour justifier une telle pratique :

— cela permet aux dialogues de provenir du même endroit, indépendamment de la place du spectateur dans le cinéma (par exemple, si un ingénieur du son latéralisait une voix entre les enceintes centrale et droite, la localisation perçue de la voix risquerait de dévier vers la droite pour les spectateurs assis en proximité de l'enceinte droite) (Theile, 2000) ;

— la cohérence audiovisuelle spatiale n'est pas indispensable, puisque l'audience perçoit une source sonore au même endroit que sa source visuelle associée même quand les deux sources ne sont pas physiquement au même endroit (Chion, 2005) : on parle alors d'*effet « ventriloque »* ;

— cela permet de gagner du temps lors du mixage ;

— spatialiser les voix peut être perçue négativement, car les spectateurs se sont bien trop habitués à voir des films « mixés au centre » : cette convention cinématographique est ainsi devenue une nouvelle référence de réalisme pour le cinéma (Chion, 2005).

Rumsey (2002) a également formulé l'hypothèse que la reproduction sonore avait fini par acquérir ses propres standards de réalisme, différents de l'écoute naturelle ;

— selon Allen (1991), les ingénieurs du son ne diffusent en général pas d'effet sonore ni de dialogue dans les « surround », car les spectateurs risqueraient de se retourner. Or, les réalisateurs ne veulent surtout pas détourner l'attention des spectateurs de l'écran.

Ces arguments seront plus largement discutés par la suite, notamment dans le chapitre 5.

En profondeur

Les ingénieurs du son se reposent en général sur des réglages de volume, de réverbération (qu'elle soit artificielle ou obtenue à partie d'une prise de son en champ diffus) et d'équalisation pour simuler les trois indices monauraux de la perception auditive de la distance (intensité, rapport champ direct/champ diffus et coloration spectrale, voir chap. 1.1.3), afin de produire une bande-son plus adaptée à l'image en termes de profondeur. Puisque ces réglages dépendent totalement de la subjectivité du mixeur, il paraît plus approprié de parler de « simulation de la profondeur » que de « cohérence audiovisuelle en profondeur ».

En élévation

Les enceintes du système 5.1 sont contenues dans un seul et même plan horizontal et ne permettent donc pas de positionner des sources dans le plan vertical.

Sons d'ambiance et musique

Les sons d'ambiance et musiques sont en général diffusés dans les enceintes avant gauche et avant droite, ainsi que dans les enceintes surround.

L'une des pratiques les plus courantes pour les sons d'ambiance est d'utiliser deux ambiances stéréophoniques décorrélées : une reproduite sur la paire d'enceintes avant gauche et avant droite, et une autre reproduite sur les 2 canaux d'enceintes surround. L'ambiance des enceintes surround est en général diffusée moins forte que l'ambiance des enceintes frontales.

Pour la musique, il n'est pas rare que les ingénieurs du son la diffusent en stéréophonie sur les enceintes avant gauche et avant droite, avec éventuellement de la réverbération dans les enceintes surround.

En résumé, les pratiques actuelles de mixage et de reproduction se caractérisent par :

— une cohérence spatiale entre son et image faible ;

— une spatialisation des sources sonores uniquement dans le plan horizontal, sans élévation ;

— une balance avant/arrière largement en faveur des enceintes frontales : dialogues et effets diffusés droit devant le spectateur, sons d'ambiance et musiques plus fortes dans les enceintes avant que dans les enceintes surround, etc. Selon Jullier (2006), « les huit

dixièmes des watts qui déferlent sur nous durant la projection proviennent de derrière l'écran ».

1.2.4 Les nouveaux systèmes de spatialisation : l'avenir du son au cinéma ?

De nouveaux formats, multipliant le nombre de canaux et d'enceintes, sont apparus récemment et pourraient bouleverser le marché du multicanal dans les années à venir.

Dolby Atmos

Le *Dolby Atmos* est un format qui a été lancé avec le film *Brave* de Mark Andrews et Brenda Chapman en 2012. Comme on peut le voir sur la Fig. 1.9, le format se caractérise par :

— l'ajout de haut-parleurs au plafond ;
— l'ajout de 2 enceintes frontales « gauche-centre » et « centre-droite » pour les écrans de plus de 12 m de large ;
— l'ajout de haut-parleurs « surround » supplémentaires (entre l'écran et les enceintes « surround » traditionnelles du 5.1) ;
— l'ajout de 2 caissons de basse en arrière de la salle ;
— le niveau de calibration ainsi que la bande passante des enceintes surround sont désormais les mêmes que pour les enceintes frontales ;
— Chaque enceinte peut-être gérée indépendamment.

Plus de 2000 salles sont aujourd'hui équipées en *Dolby Atmos* dans le monde.

Auro-3D

Auro Technologies propose deux formats : *Auro-3D* 11.1 et 13.1. La configuration 11.1 est en fait un 5.1 classique, auquel s'ajoute un autre 5.0 en hauteur ainsi qu'un canal « VOG » (« Voice of God », aussi appelé « enceinte douche ») au dessus du public (voir Fig. 1.10). L'ajout d'un canal surround central et son équivalent zénithal permet d'obtenir la configuration 13.1. À l'été 2015, Auro Technologies annonçait 80 salles équipées en *Auro-3D* dans le monde, et 300 projets de salles confirmés.

22.2 de la NHK

La NHK (compagnie nationale de diffusion au Japon) propose un format 22.2, avec trois couches de haut-parleurs (voir Fig. 1.11) :

— une couche zénithale à 9 canaux ;
— une couche médiane (à hauteur de tête) à 10 canaux ;
— une couche basse à 3 canaux située sous l'écran.

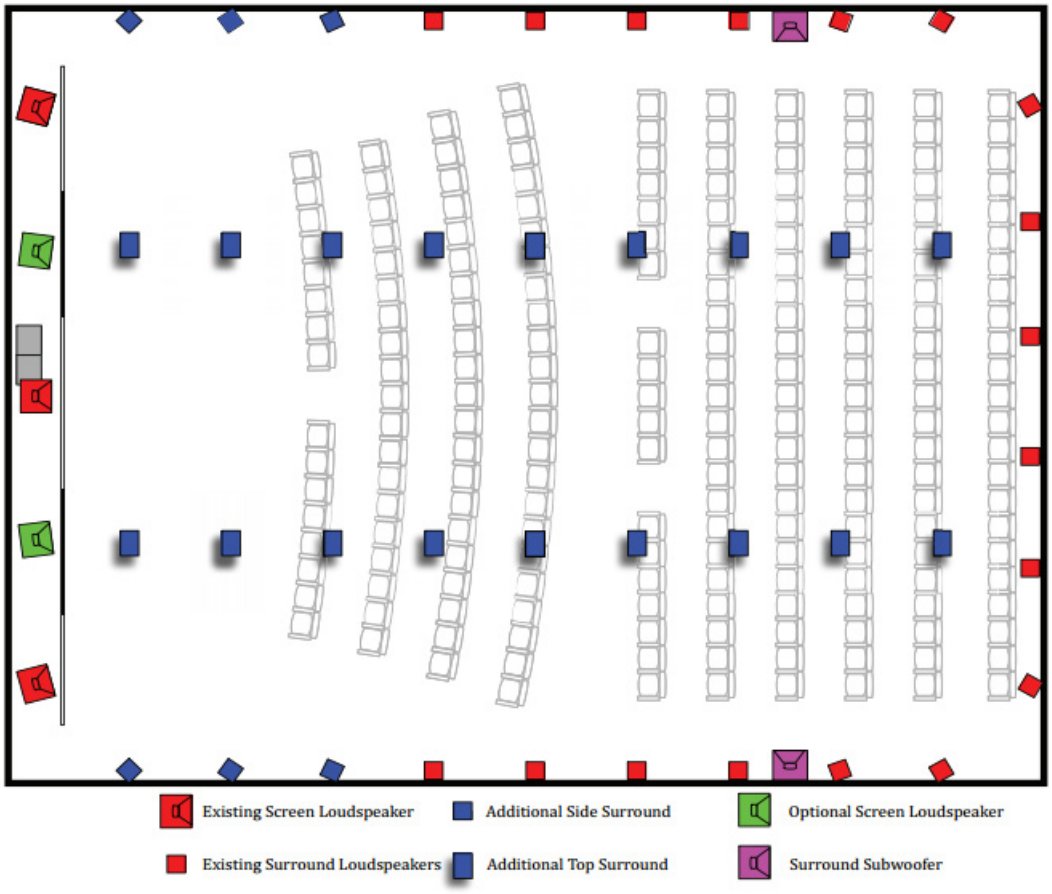


FIGURE 1.9 – Configuration Dolby Atmos et comparaison avec un système 5.1 classique (Dolby, 2012). Le rectangle gris à côté de l'enceinte centrale représente le caisson de basse déjà présent dans le système 5.1 classique.

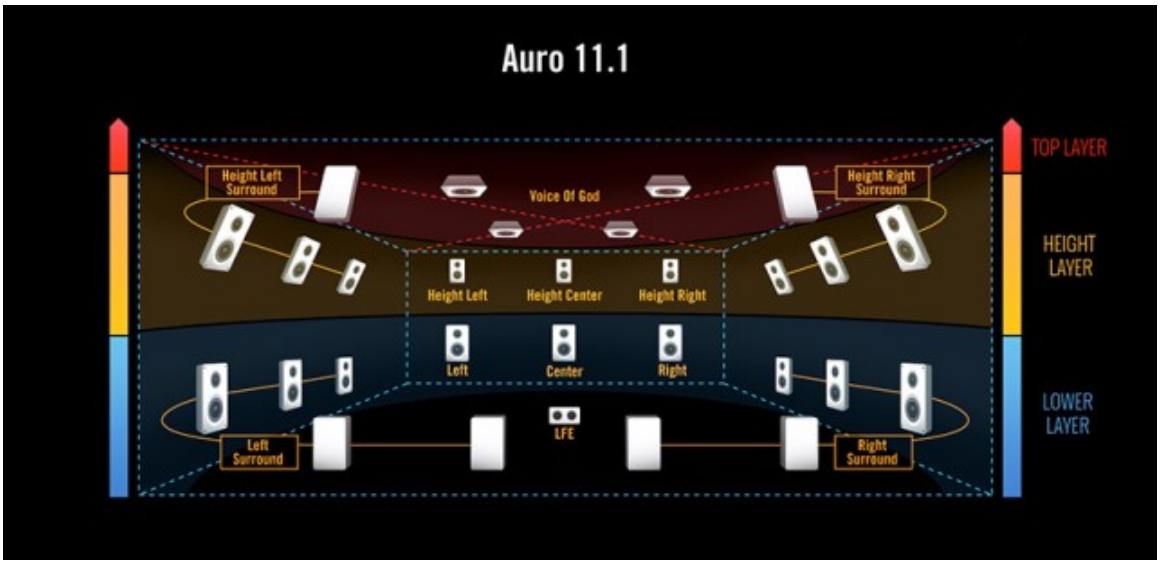


FIGURE 1.10 – Configuration Auro 3D 11.1 (Auro-3D, 2006).

Dolby Atmos, *Auro-3D* et le 22.2 partagent les mêmes caractéristiques que les systèmes multicanaux tels que le 5.1 ou le 7.1 en termes d'indices de la localisation correctement rendus et non rendus. A la différence du 5.1, il s'agit cependant de systèmes 2-D, car les sources

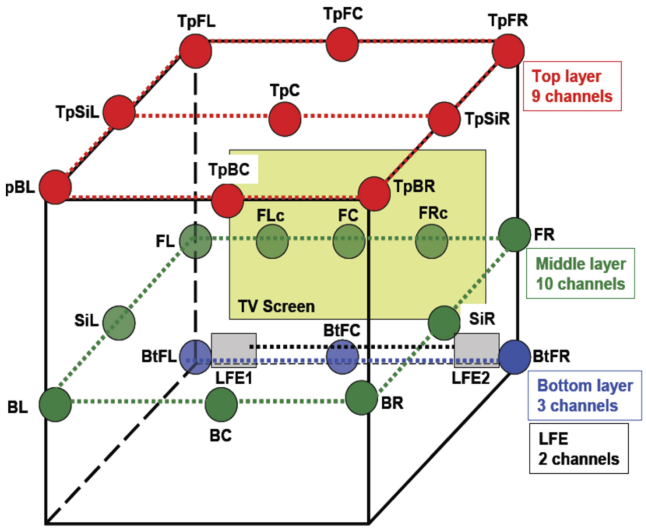


FIGURE 1.11 – Configuration 22.2. D'après Hamasaki *et al.* (2004).

peuvent à la fois évoluer en azimut et en élévation. Par contre, la distance est toujours fixée aux positions des enceintes.

WFS

Plusieurs études ont envisagé la WFS comme système de reproduction pour le cinéma (Sporer, 2004), et plus particulièrement pour le cinéma 3D (André *et al.*, 2014; Moulin, 2015). En février 2003, le premier cinéma équipé en WFS (192 haut-parleurs) a été inauguré à Ilmenau, en Allemagne (voir Fig. 1.12). En juillet 2004, c'est au tour d'Hollywood d'équiper un de ses studios avec un système WFS comprenant 304 haut-parleurs.

FIGURE 1.12 – La salle de cinéma équipée en WFS à Ilmenau (Allemagne).

Selon le principe de Huygens, chaque point d'un front d'onde généré par une source primaire peut être considéré comme une source secondaire émettant des ondes secondaires. L'idée de la WFS, proposée par Berkhout (1988; 1993), est de remplacer les sources secondaires par des haut-parleurs et de superposer les ondes secondaires générées par ces haut-parleurs pour obtenir une approximation de ce que serait effectivement l'onde sonore si elle était générée

par la source primaire (voir Fig. 1.13). Il s'agit donc d'une **restitution physique du champ sonore**, par opposition aux systèmes que nous avons présentés précédemment, qui étaient des systèmes de **restitution psycho-acoustique du champ sonore**, principalement basés sur des **illusions auditives** (images fantômes pour placer des sources dans la latéralité, simulation des indices monauraux de perception de la distance pour donner une impression de profondeur, etc.).

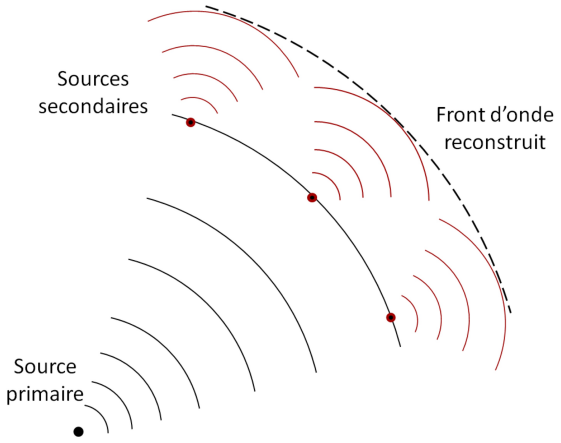


FIGURE 1.13 – Illustration du principe de Huygens.

La WFS est basée sur des approximations de l'intégrale de Kirchhoff-Helmholtz, permettant d'exprimer un champ acoustique sur le sous-espace Ω_2 (la zone d'écoute), engendré par des sources primaires présentes dans un sous-espace Ω_1, à l'aide de sources placées sur la frontière séparant les deux sous-espaces. La pression acoustique $p(\vec{r})$, en tout point $\vec{r}$ à l'intérieur de Ω_2, est définie grâce à la pression p_0 et au gradient de pression $\vec{\nabla}_{p_0}$ sur la surface de séparation entre Ω_1 et Ω_2 (notée $\partial\Omega_0$) par l'équation intégrale

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

Dans cette intégrale, le vecteur $\vec{n}$ représente le vecteur unitaire normal à la surface $\partial\Omega_0$ et extérieure au domaine Ω_2, et le vecteur $\vec{R}$ 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 et un dipôle, ce qui revient à décomposer respectivement le signal acoustique en pression et en vitesse acoustique. Contrairement au Principe de Huygens, qui 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 (Moulin, 2015).

Plusieurs précautions ou simplifications doivent être entreprises pour l'implémentation

pratique du système :

- les réflexions de la salle d'écoute peuvent générer d'importantes erreurs de localisation. Une équalisation des haut-parleurs peut permettre de réduire efficacement l'influence de la salle (Corteel et Nicol, 2003) ;
- la reproduction est souvent limitée à une ligne de haut-parleurs, ce qui entraîne la limitation du rendu sonore spatialisé au plan horizontal, et un rayonnement du champ sonore à symétrie cylindrique et non sphérique. L'amplitude est alors réduite de 3 dB par doublement de la distance ;
- cette ligne de haut-parleurs est souvent limitée en longueur, ce qui provoque une limitation de la zone d'écoute dans laquelle les sources peuvent être restituées correctement. De plus, la réduction à un segment de haut-parleurs génère des phénomènes de diffraction : le front d'onde est correctement synthétisé, mais des fronts d'onde secondaires retardés sont également générés, qui peuvent soit colorer le front d'onde primaire soit être perçus comme des échos. Pour limiter les phénomènes de diffraction, des gains plus faibles peuvent être appliqués sur les enceintes placées aux extrémités (Vogel, 1993), mais cette solution a l'effet indésirable de réduire encore plus la zone d'écoute (Boone *et al.*, 1995) ;
- enfin, la ligne de haut-parleurs ne pourra pas être considérée comme une ligne continue de sources secondaires puisqu'elle est composée de plusieurs haut-parleurs. Ces haut-parleurs peuvent avoir des tailles différentes, être collés les uns aux autres (« *high-density WFS* ») ou au contraire plus espacés (« *low-density WFS* »). Cette discrétisation de la ligne va provoquer un phénomène de repliement spatial (« *spatial aliasing* ») : Soit f_{Nyq} la fréquence d'aliasing définie par :

  $$f_{Nyq} = c/2\Delta x$$

  avec c la célérité du son dans l'air et Δx la distance entre chaque haut-parleur.

  Tant que la fréquence est inférieure à la fréquence d'aliasing, la reproduction reste stable dans toute la zone d'écoute. Cela suppose un espacement, peu réaliste, d'au plus 8.5 mm entre chaque enceinte pour obtenir une reproduction correcte jusqu'à 20 kHz.

  Avec une distance entre les haut-parleurs de 11 cm (soit une fréquence d'aliasing légèrement supérieure à 1500 Hz), Start (1997) a montré que le repliement spatial n'avait pas d'impact critique en termes de localisation sonore par rapport à une diffusion sur enceinte monophonique (les stimuli utilisés étaient des bruits blancs plus ou moins larges-bandes). Cependant, il constate que des artefacts audibles tels que dégradation du timbre peuvent être perçus.

Performances de localisation

Plusieurs études ont montré que l'acuité et la précision de localisation en azimut avec la WFS étaient comparables à celles de la « vraie vie » (Rébillat *et al.*, 2008, 2012; Verheijen, 1998). Avec un espacement de 22 cm entre enceintes, Verheijen (1998), par exemple, a obtenu une erreur de localisation moyenne de 3.2° (variance de 1.4°) avec le système WFS contre 2.6° (variance de 1°) avec une source réelle.

Rohr *et al.* (2013) ont proposé une « version 3D » de la WFS (« *low-density* ») intégrant la dimension verticale, et ont également obtenu de bonnes performances de localisation en élévation.

Ainsi, pour des fréquences inférieures à la fréquence d'aliasing, la WFS est capable de reproduire correctement les ITD, ILD et indices de la profondeur. Les indices spectraux ne sont par contre pas reproduits et demeurent ceux des enceintes. Les systèmes WFS se déclinent en version 2-D (azimut et profondeur) ou alors, comme nous venons de le voir, en version 3-D (azimut, profondeur et élévation).

1.2.5 Conclusion

Par rapport au système 5.1 classique, les améliorations que proposent les nouveaux systèmes de spatialisation peuvent se résumer en deux points principaux :

— **une meilleure cohérence audiovisuelle spatiale** :
  - dans le plan horizontal : Dolby Atmos, 22.2, WFS ;
  - dans le plan vertical : Dolby Atmos, 22.2, Auro-3D, « version 3D » de la WFS ;
  - dans la profondeur : WFS.

— **une plus grande importance accordée au « surround »** :
  - discrétisation des canaux, permettant de gérer individuellement chaque enceinte : Dolby Atmos, 22.2, WFS ;
  - ajout d'enceintes surround en proximité de l'écran : Dolby Atmos ;
  - ajout d'enceintes surround zénithales : Dolby Atmos, 22.2, Auro-3D ;
  - égalité de niveau et de bande passante entre enceintes surround et enceintes frontales : Dolby Atmos, 22.2, Auro-3D ;
  - ajout de caissons de basse dédiés : Dolby Atmos.

Ces nouveaux systèmes remettent donc en cause les pratiques actuelles de mixages (voir chap. 1.2.3) puisqu'ils incitent à ne plus forcément « mixer au centre », à ne pas limiter la diffusion du son dans le plan horizontal, et à accorder beaucoup plus d'importance au surround.

Le but de notre étude est de déterminer si la stéréoscopie influence la perception du son au cinéma. Nous tenterons de mettre en évidence cette influence dans des cas de figures simples et proches des réalités actuelles. Ainsi, nous utiliserons comme dispositifs pour nos

expériences le 5.1 traditionnel (expériences I et II), ainsi qu'un 5.1 avec ajout de deux enceintes frontales (expérience V), comme dans les formats Cinerama (1952), Cinemascope (1953) et SDDS (1993). L'expérience III ne porte que sur l'image et ne comportera donc aucun son. Quant à l'expérience IV, elle n'utilise aucun format prédefini et son dispositif sera présenté ultérieurement.

Chapitre 2

L'image

Sommaire

Ce chapitre se focalise sur la perception visuelle et sur l'image stéréoscopique. Dans un premier temps, nous présentons les mécanismes et performances de la localisation visuelle. Nous décrivons ensuite les principes de la stéréoscopie, de la captation à la diffusion. Enfin, nous terminons par un bref historique du cinéma en relief.

2.1 La localisation visuelle

Pour un oeil, le champ de vision horizontal est limité à 100° vers l'extérieur (côté tempe), et à 60° vers l'intérieur (côté nez), soit une couverture totale de 180° horizontalement quand on somme la contribution des deux yeux, et une zone de recouvrement d'environ 120° dans laquelle la vision est binoculaire (Spector, 1990).

Quant au champ de vision vertical, il est limité à 40° vers le haut et à 70° vers le bas, soit un total de 110°.

2.1.1 Perception visuelle de la direction

Pour une scène visuelle donnée, le cerveau reçoit deux images rétiniennes « œil gauche » et « œil droit » différentes. Sur ces deux images, l'objet est perçu dans une direction différente

(la direction « oculocentrique »). Pourtant, une fois les deux images fusionnées par le cerveau, un observateur perçoit bien l'objet dans une seule et unique direction (la direction « égocentrique »). Le cerveau déduit donc une direction égocentrique à partir de deux directions oculocentriques.

Selon Hering (1942), l'homme percevrait le monde comme s'il avait un troisième œil entre les deux yeux, l'œil « cyclopéen ». Cependant, cette hypothèse a été remise en question il y a quelques années par Erkelens et van Ee (2002).

2.1.2 Perception visuelle de la profondeur

Les indices de la localisation visuelle en profondeur peuvent être divisés en deux familles (Jouhaneau, 2012) :

— les indices exploitables par un seul œil : ce sont les indices « monoculaires » ;
— les indices basés sur une différence entre les images rétiniennes de chaque œil : ce sont les indices « binoculaires ».

La superposition de deux indices peut renforcer la sensation de profondeur ou au contraire la perturber s'ils sont contradictoires. L'importance de chaque indice varie essentiellement avec la distance de l'objet : la disparité binoculaire, par exemple, est très importante en champ proche mais devient négligeable en champ lointain.

Indices monoculaires

Les indices monoculaires se divisent en deux catégories : les indices physiologiques et les indices psychologiques acquis par apprentissage.

Indices physiologiques

— L'accommodation : le cristallin des yeux se déforme pour effectuer la mise au point sur un objet (qu'il soit proche ou lointain) et permettre ainsi une plus grande netteté de vision. Cette déformation spécifique peut être interprétée par le cerveau comme un indice permettant une estimation absolue de la distance d'un objet visuel ;
— La parallaxe : lors d'un déplacement latéral entre l'objet et l'observateur, l'observateur effectue généralement un suivi des objets par rotation de son axe de vision, donc de ses yeux et, éventuellement, de sa tête. Le changement angulaire est de faible amplitude pour un objet éloigné, et plus conséquente pour un objet proche. Il s'agit donc d'un indice de distance relative.

Indices psychologiques

— Effet de taille : la taille des objets diminue avec la distance. Cet indice 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 ;

— Interposition (ou occlusion) : un objet partiellement masqué par un autre se trouve forcément plus loin que l'objet masquant. L'occlusion est donc un indice de distance relative ;

— Hauteur dans le champ visuel : plus un objet s'éloigne, plus il se rapproche de la ligne d'horizon. Cet indice permet d'estimer la distance de manière absolue ;

— Ombres : le fait de savoir que la lumière solaire ou l'éclairage d'une pièce vient généralement d'en haut aura une influence sur la perception des formes et leur position relative ;

Ces quatre indices sont présentés dans la Fig. 2.1.

• **Effet de taille** : la chaise de droite étant plus petite que celle de gauche, on en conclut qu'elle est plus éloignée.

• **Effet d'interposition** : la boîte, masquée par le pied avant gauche de la chaise et masquant le pied arrière est évaluée comme étant située entre les deux.

• **Effet de hauteur** : la position relative des pieds sur le sol indique une nette différence d'éloignement des deux chaises.

• **Effet d'ombre** : la portion d'ombre verticale de la chaise de droite conduit à la conclusion qu'elle est plus proche du mur que celle de gauche.

FIGURE 2.1 – llustration des effets de taille, d'interposition, de hauteur dans le champ visuel et d'ombre sur la perception de distance apparente d'un objet. D'après Jouhaneau (2012).

— La perspective linéaire : souvent utilisée en peinture pour créer une impression de profondeur (lignes au sol, orientation des murs), la perspective linéaire se traduit par la présence de lignes convergentes vers un (ou deux) points de fuite (voir Fig. 2.2). Il s'agit d'un indice de distance relative ;

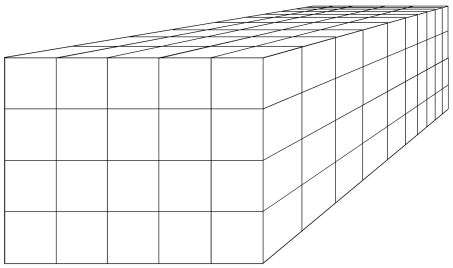


FIGURE 2.2 – llustration de la perspective linéaire. D'après Moulin (2015).

— La perspective chromatique : la surface uniforme d'un objet peut changer de couleur

avec la distance. Le ciel par exemple est bleu au-dessus de notre tête mais devient presque blanc à l'horizon (voir Fig. 2.3). Il s'agit d'un indice de distance relative ;

— La perspective de texture : plus un objet est proche, plus il sera facile de distinguer la texture de sa surface. Lorsqu'un observateur regarde la mer par exemple, il distingue clairement les vagues au premier plan. Par contre, en champ lointain, la texture perçue de la mer est totalement uniforme (voir Fig. 2.3). Il s'agit d'un indice de distance relative ;

FIGURE 2.3 – lllustration de la perspective chromatique (au loin, le ciel devient presque blanc) et de la perspective de texture (les vagues sont de moins en moins distinctes avec la distance).

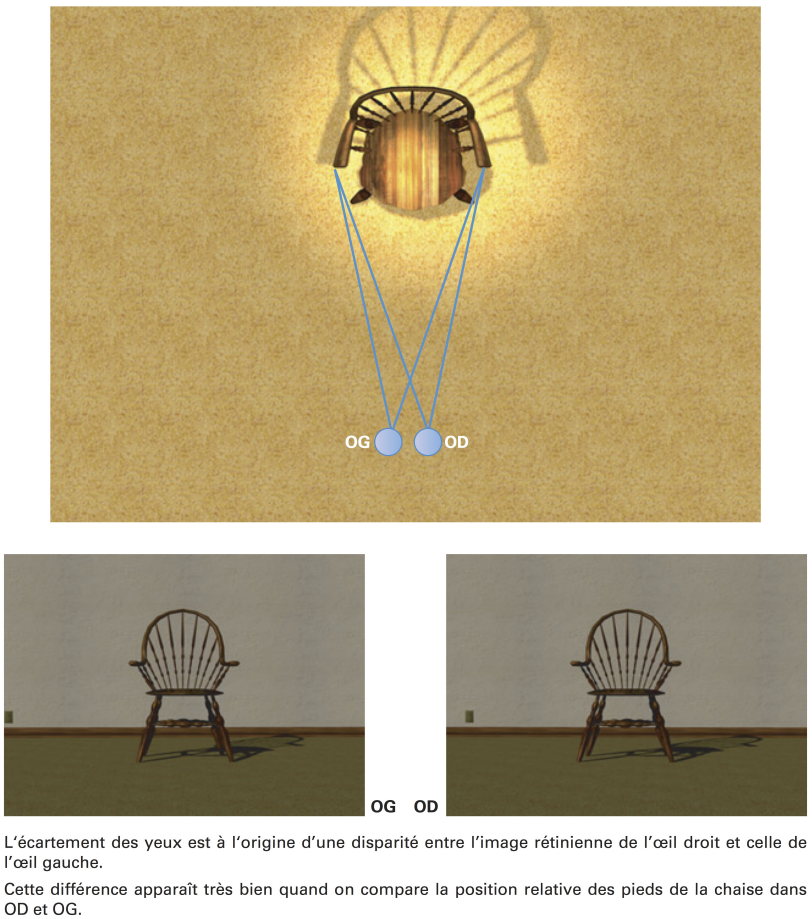
— La perspective de diffusion : un objet lointain est perçu flou et moins contrasté qu'on objet proche à cause de l'humidité, de la pollution atmosphérique, du brouillard, etc. Cet indice permet d'estimer la position absolue d'un objet, pour des distances importantes.

Indices binoculaires

Les indices binoculaires sont des informations de profondeur basées sur la différence entre les images rétiniennes de chaque œil. Cette différence est causée par le décalage entre les yeux d'un observateur, appelé écart interoculaire (en moyenne de 62 mm chez la femme et 65 mm chez l'homme adulte (Dodgson, 2004)). Ces images sont comparées par le cerveau puis fusionnées en une seule et même image spatialisée.

— La disparité binoculaire : il s'agit du décalage azimutal entre les deux images rétiniennes d'un même objet (voir Fig. 2.4). Plus le décalage entre les images sera important, plus l'objet sera perçu proche. La disparité binoculaire permet de définir la position relative des objets devant et derrière le point de fixation ;

— La convergence : lorsqu'un observateur fixe un objet, il modifie la mise au point de chaque cristallin (l'accommodation, voir ci-dessus) mais fait aussi converger les axes optiques de ses yeux vers le point de fixation. Plus l'objet est proche, plus la convergence des yeux est importante. Cet indice permet ainsi une estimation absolue de la distance d'objets visuels.



L'écartement des yeux est à l'origine d'une disparité entre l'image rétinienne de l'œil droit et celle de l'œil gauche.

Cette différence apparaît très bien quand on compare la position relative des pieds de la chaise dans OD et OG.

FIGURE 2.4 – Disparité binoculaire : Exemple visuel du décalage azimutal de l'image rétinienne. D'après Jouhaneau (2012).

2.1.3 Performances

L'acuité visuelle monoculaire

L'acuité visuelle monoculaire est très précise dans le champ de vision central : Howard (1982) a montré que deux points pouvaient y être vu séparément même si l'écart angulaire entre eux était de l'ordre d'une minute d'arc (soit 1/60 de degré). Cavonius et Robbins (1973) ont également obtenu des seuils et écart-types de cet ordre-là.

L'acuité visuelle binoculaire

L'acuité visuelle binoculaire (aussi appelée « acuité stéréoscopique ») est définie par la disparité binoculaire à partir de laquelle un être humain est capable de distinguer une diffé-

rence de profondeur entre deux objets. Selon les individus, le seuil d'acuité peut aller de 30 à 2 secondes d'arc (Coutant et Westheimer, 1993), 2 secondes signifiant que l'individu peut détecter une différence de profondeur de 4 mm à une distance de 5 mètres.

2.2 La stéréoscopie

2.2.1 Les systèmes de captation stéréoscopique

Un système stéréoscopique nécessite au moins deux images planes (une pour chaque oeil) pour pouvoir reproduire une perception du relief : la scène doit donc être captée depuis 2 angles de vue différents, avec 2 caméras.

Les caméras doivent être placées côte à côte, avec une distance interaxiale (distance entre les deux lentilles) proche de la distance interoculaire moyenne. Si les systèmes optiques sont trop encombrants, les deux caméras peuvent être placées perpendiculairement, comme illustré dans la Fig. 2.5. Grâce à un miroir semi-réfléchissant, la caméra dans l'axe horizontal capte la lumière réfractée par le miroir tandis que la caméra dans l'axe verticale capte la lumière réfléchie.

Les deux caméras utilisées dans le cadre de la présente étude sont présentées dans la Fig. 2.6.

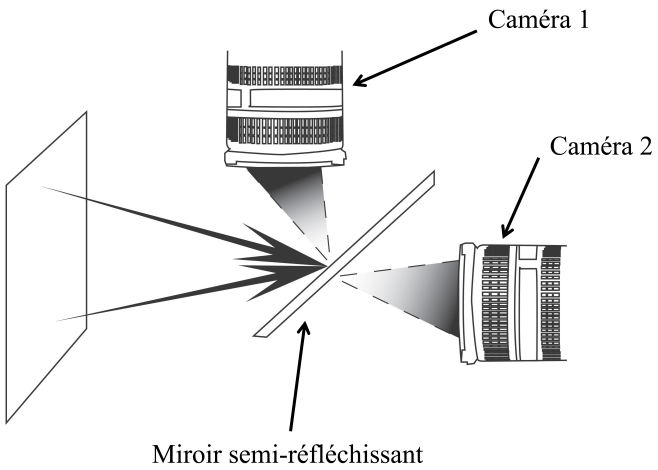


FIGURE 2.5 – Système à 2 caméras avec miroir semi-réfléchissant.

FIGURE 2.6 – Les deux caméras utilisées dans le cadre de la présente étude.
A gauche : Caméra Panasonic AG-3DP1 avec lentilles côte à côte.
A droite : Sinacam 3D avec miroir semi-réfléchissant.

2.2.2 Perception d'images stéréoscopiques

Le but d'un système stéréoscopique est de reproduire une perception du relief en présentant à chaque œil deux images planes différentes.

Soit un objet 3D, assimilé à un point, divisé en deux objets 2D « gauche » et « droit » (c'est-à-dire destinés à être respectivement vu par l'œil gauche et l'œil droit) projetés sur un écran. Le décalage entre l'objet « gauche » et l'objet « droit » est appelé disparité stéréoscopique. La position relative de ces deux objets 2D sur l'écran va déterminer la position de l'objet 3D : il suffit de tracer une ligne entre l'œil gauche du spectateur et l'objet « gauche » et une autre ligne entre l'œil droit du spectateur et l'objet « droit ». Le croisement de ces deux lignes détermine la position de l'objet 3D.

— s'il n'y a pas de décalage entre l'objet « gauche » et l'objet « droit », on parle de disparité nulle et l'objet 3D résultant est perçu au niveau de l'écran, sans relief ;

— si, sur l'écran, la position de l'objet « droit » se trouve à droite de la position de l'objet « gauche », on parle de disparité positive et l'objet 3D est perçu derrière l'écran. Plus la disparité est importante, plus l'objet 3D est perçu loin derrière l'écran. Si la disparité atteint la distance interoculaire, alors les lignes sont parallèles et l'objet 3D est perçu à l'infini. Il faudra cependant veiller à ce que la disparité ne dépasse pas la distance interoculaire : dans ce cas, les lignes ne se croisent plus et l'observateur perçoit deux images différentes (on parle alors de divergence).

— si, sur l'écran, la position de l'objet « droit » se trouve à gauche de la position de l'objet « gauche », on parle de disparité négative et l'objet 3D est perçu devant l'écran. On parle alors d'images en jaillissement. Plus la disparité est importante, plus l'objet 3D se rapproche du spectateur.

Les différents types de disparité sont illustrés sur la Fig. 2.7.

Dans le chapitre 2.1.2, nous avons vu qu'un observateur, lorsqu'il fixe un objet :

— modifie la mise au point (et donc la distance focale) de chaque cristallin pour en avoir une image nette : l'*accommodation*.

— fait converger les axes optiques de ses yeux vers le point de fixation : la *convergence*.

En vision naturelle, la distance de convergence et la distance focale sont égales, et accommodation et vergence sont neuralement connectées. Cependant, lorsqu'un observateur regarde un contenu stéréoscopique, la distance focale est égale à la distance entre les yeux et l'écran du système de restitution, alors que la distance de convergence est égale à la distance entre les yeux et l'objet virtuel regardé. Pour voir l'objet clairement et sans diplopie (vision double), l'observateur doit donc accommoder à une distance différente de celle à laquelle il converge, ce qui nécessite de neutraliser la connexion neuronale entre vergence et accommodation.

Ce conflit accomodation/vergence peut dans une certaine mesure être traité par le système perceptif, mais il peut également générer de la fatigue visuelle et de l'inconfort si les disparités

sont trop importantes. Il est donc recommandé de respecter une zone de confort, c'est-à-dire de placer les objets de la scène à restituer dans un intervalle limité de profondeur devant et derrière l'écran. Cette zone de confort se situe principalement derrière l'écran (Mendiburu, 2009), ce qui explique pourquoi, selon Kunka et Kostek (2013), les productions professionnelles en 3D exploitent principalement des disparités positives.

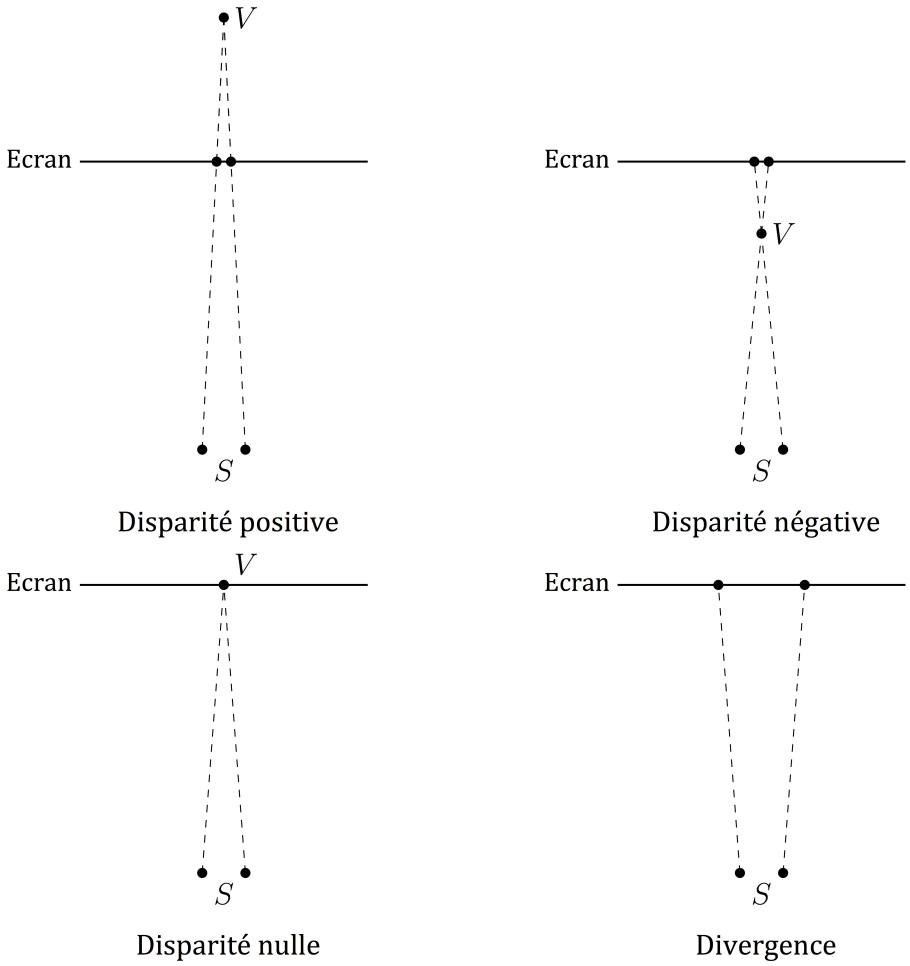


FIGURE 2.7 – Un spectateur S (les deux points représentent ses deux yeux) perçoit un objet 3D plus ou moins devant ou derrière l'écran selon la disparité stéréoscopique. Si la disparité est trop grande (divergence), le spectateur perçoit deux images différentes. D'après André (2013).

2.2.3 Les systèmes de restitution stéréoscopique

La Fig. 2.8 présente le principe du premier dispositif de vision stéréoscopique, conçu par Wheatstone en 1838. D'autres procédés moins encombrants furent proposés par la suite, qui peuvent être divisés en deux catégories : les systèmes passifs et les systèmes actifs.

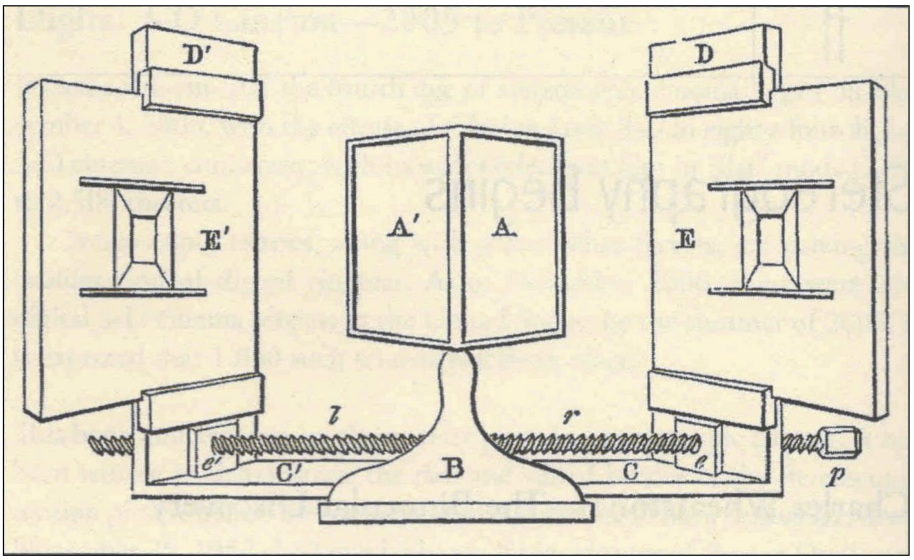


FIGURE 2.8 – Stéréoscope de Charles Wheatstone. Les photographies sont placées en E et E'. Deux miroirs sont placés à angle droit en A et A'. En s'approchant suffisamment des miroirs, l'œil droit ne voit plus que l'image E et l'œil gauche l'image E'.

Systèmes passifs

— Anaglyphes : Les anaglyphes reposent sur la présentation de deux images colorées figurant sur un même support. Les deux images, en général colorées l'une en cyan et l'autre en rouge, sont vues à travers des lunettes portant deux filtres également de couleurs cyan et rouge (voir Fig. 2.9). Chaque œil ne voit ainsi que l'image qui lui est destinée. Il en résulte une restitution médiocre et non homogène d'un œil à l'autre de la couleur et de la luminosité, ce qui cantonne ce système à la diffusion d'images principalement monochromes.

— Filtres polarisants : Deux projecteurs, l'un fonctionnant avec une lumière polarisée verticalement, l'autre avec une lumière polarisée horizontalement, envoient simultanément les deux images sur l'écran. Le spectateur, qui regarde l'écran avec une paire de lunettes dont les verres sont orthogonalement polarisés, ne verra que l'image à polarisation horizontale sur un œil, et l'image à polarisation verticale sur l'autre. Il en résulte une restitution bien meilleure de la couleur. La luminosité est également mieux restituée mais les pertes restent importantes à cause du filtrage. 75% des salles en Amérique du Nord seraient équipées avec cette technologie (Jones, 2009).

Systèmes actifs

— Images successives : Dans ce procédé plus récent, les images ne sont plus présentées simultanément par deux projecteurs, mais successivement par un seul et même projecteur, en alternant rapidement (jusqu'à 48 i/s) les images destinées à l'œil gauche et à l'œil droit à l'aide d'un obturateur placé devant chaque œil. Les lunettes portant

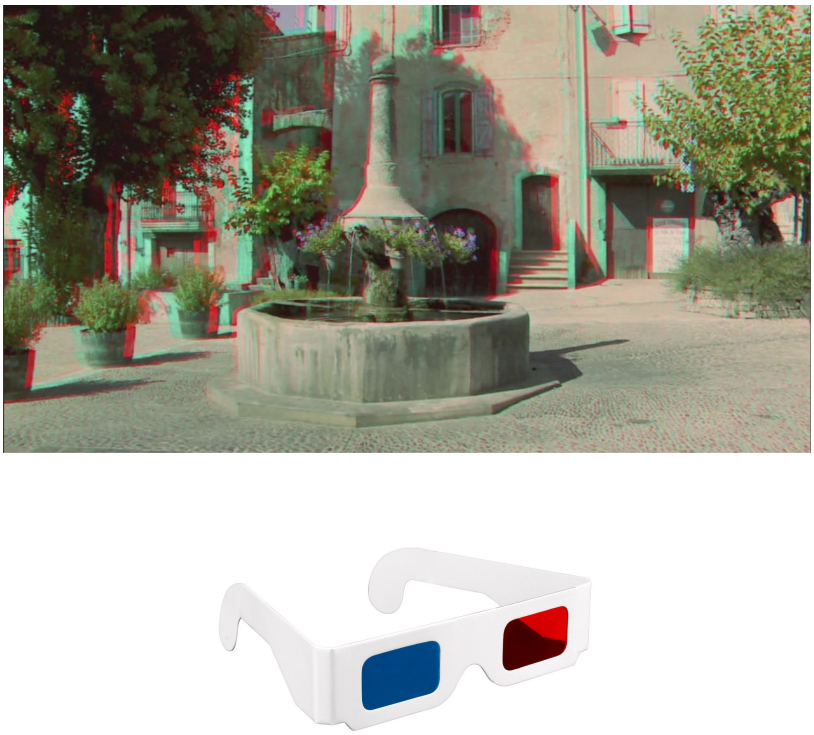


FIGURE 2.9 – Vue anaglyphique (d'après Gambier (2010)) et lunettes 3D anaglyphiques.

les obturateurs sont synchronisées avec la projection, de telle sorte que chaque œil ne reçoive que les images qui lui sont destinées. On parle alors de lunettes actives. Malgré une perte de luminosité (de 40 à 70%) plus importante que pour les systèmes à filtres polarisants (Woods, 2001), cette technologie est très largement employée en Europe dans les cinémas et sur certains téléviseurs (Jones, 2009). Les expériences de la présente étude ont été réalisées avec des lunettes actives (voir Fig. 2.10).

FIGURE 2.10 – Lunettes 3D actives Epson ELPGS01 (utilisées dans les expériences I, IV et V) et Eyes3Shut Purple Two (utilisées dans les expériences II et III).

Il existe également des technologies dites autostéréoscopiques, qui proposent une restitution en relief sans lunettes. Cependant, ces technologies imposent au spectateur d'être placé à un endroit bien précis en face de l'écran, ce qui freine considérablement leur développement.

2.3 Le cinéma 3D : un bref historique

Les premières expériences

Dès 1903, les frères Lumière présentent plusieurs court-métrages stéréoscopiques, mais pour un seul spectateur à la fois. La première projection publique en relief aurait eu lieu le 10 juin 1915 à New York, avec trois courtes séquences projetées en anaglyphe. De nombreuses productions sont tournées en relief dans les années 20, dont *the Power of Love* (1922) de Nat G. Deverich et Harry K. Fairall, considéré comme le premier long métrage en relief (en anaglyphe également). Cependant, l'arrivée du cinéma « parlant » en 1927 éclipse le relief et le relègue aux films expérimentaux ou érotiques.

L'engouement des années 50

La sortie de *Bwana Devil* d'Arch Oboler, le 26 novembre 1952, marque le début d'un âge d'or fugitif pour le cinéma en relief. Suivront notamment *House of Wax*, d'André de Toth (1953), première grosse production en relief (il s'agit également du premier film utilisant un son stéréophonique) et *Dial M for Murder* d'Alfred Hitchcock (1954). Dans la période 1952-1955, une cinquantaine de films seront tournés en relief, mais la production en 3D finira par s'essouffler, à cause de problèmes techniques (défauts de synchronisme entre les images « œil droit » et « œil gauche », manque de lumière pour l'image, etc.), artistiques (qualité médiocre des films, utilisation abusive des effets de jaillissement, etc.) et de confort pour les spectateurs (port des lunettes obligatoire).

La « traversée du désert » (années 70-80)

Pendant les années 70 et 80, la production en relief reste marginale, limitée aux films pornographiques, « gore » et expérimentaux, ou à quelques tentatives de films d'horreur tels que *Jaws 3-D* de Joe Alves, *Amityville 3-D* de Richard Fleischer ou *Friday the 13th Part III* de Steve Miner.

La projection numérique et *Avatar*

L'arrivée de la projection numérique, en 1999, va permettre de régler le problème de synchronisme des deux images « œil droit » et « œil gauche ».

Après quelques films comme *Ghosts of the Abyss* de James Cameron (2003), *The Polar Express* de Robert Zemeckis (2004) ou *Chicken Little* de Mark Dindal (2005), la sortie d'*Avatar* de James Cameron, en 2009, plus gros succès de l'histoire du cinéma après seulement six semaines d'exploitation, va susciter un engouement inédit pour le cinéma 3D : la France comptabilise 14,8 millions d'entrées en France, dont 11,5 en relief (Djian, 2013). Les sociétés américaines Dreamworks et Disney annoncent produire tous leurs films en 3D à partir de 2009

et exigent même dans un premier temps que les films soient uniquement projetés dans leur version 3D. Cependant, dès 2011, on constate un déclin de l'intérêt du public pour la 3D, aussi bien en Amérique qu'en Europe. « Mais comment voulez-vous gagner la confiance du public en gonflant en 3D, n'importe comment, à la va-vite, des ratages comme *Le choc des Titans* ? » s'irrite Jocelyn Bouyssy, directeur général du groupe CGR (3ème circuit cinématographique Français avec plus de 400 salles) (Djian, 2013). Pour Frédéric Monnereau, chef des ventes de la 20th Century Fox, le déclin s'explique par le fait que les films en 3D ne sont « ni pensés ni tournés en relief » (Djian, 2013).

L'exemple récent de *Gravity* (2013) d'Alfonso Cuarón montre que le succès peut encore être au rendez-vous pour le cinéma en relief : aux Etats-Unis, 80% des recettes du film sont venues des projections en 3D, soit plus que pour *Avatar* (72%). Cet exemple montre qu'« il faut être rigoureux et choisir les bons films à diffuser dans ce format. Il faut que le relief apporte quelque chose à l'histoire, qu'il ne soit pas gratuit. Alors le film présenté aura une chance de devenir un phénomène culturel » (Dan Fellman, chargé de la distribution des productions de la Warner Bros aux Etats-Unis (Première, 2013)). *Gravity* a également été acclamé pour sa bande-son audacieuse et novatrice : pas de son dans l'espace, et une spatialisation permanente des voix, des effets sonores et de la musique tout autour des spectateurs. Pour le réalisateur Alfonso Cuarón, la bande-son du film participe autant que la 3D à l'immersion du spectateur, ce qui lui a d'ailleurs valu de remporter les Oscars du meilleur montage son, du meilleur mixage son et de la meilleure musique. Cet exemple montre que le son a un rôle important à jouer pour convaincre le public d'aller découvrir des films au cinéma en relief, ce qui confirme le besoin d'une compréhension accrue de l'influence de la stéréoscopie sur la perception du son.

Chapitre 3

L'image et le son

Sommaire

Dans ce chapitre, nous abordons différents phénomènes d'interactions entre le son et l'image susceptibles d'influencer l'expérience des spectateurs au cinéma. Dans un premier temps, nous présentons une synthèse des diverses études ayant été menées sur l'*effet ventriloque*. Lorsque l'on présente à un sujet des stimuli audio et visuel temporellement coïncidents mais spatialement disparates, les sujets perçoivent parfois le stimulus sonore au même endroit que le stimulus visuel. On appelle ce phénomène l'*effet ventriloque* car il rappelle l'illusion créée par le ventriloque lorsque sa voix semble plutôt provenir de sa marionnette que de sa propre bouche. Dans un deuxième temps, nous proposons un état de l'art des études ayant été conduites sur la perception du son lié à l'image en 3D-s.

3.1 L'effet ventriloque

L'effet d'une disparité spatiale entre deux stimuli sonore et visuel associés peut être étudié de deux manières (Bertelson et Radeau, 1981) :

— soit à l'aide d'une tâche de localisation, qui permet d'observer des biais intersensoriels (*cross-modal bias* en anglais) ;

— soit à l'aide d'une tâche de discrimination, qui permet d'observer l'*effet ventriloque* à proprement parler (on parle également dans ce cas de *fusion perceptive*).

3.1.1 Biais intersensoriels

Dans une tâche de localisation, les sujets doivent indiquer la direction d'où provient le stimulus sonore. Le jugement de localisation est souvent biaisé et la position du stimulus sonore est attirée par le stimulus visuel associé (Alais et Burr, 2004; Battaglia *et al.*, 2003; Bertelson et Aschersleben, 1998; Bertelson et Radeau, 1981; Hairston *et al.*, 2003; Wallace *et al.*, 2004; Bermant et Welch, 1976; Pick *et al.*, 1969; Radeau, 1974; Radeau et Bertelson, 1976; Warren, 1979; Weerts et Thurlow, 1971). Par exemple, avec une disparité de 7° entre des stimuli sonore et visuel dans le plan horizontal, Bertelson et Radeau (1981) ont observé que la localisation de la source sonore par les sujets étaient décalée de 4° vers le stimulus visuel. Ils ont également observé un décalage d'environ 6.3° pour une différence de 15° entre stimuli sonore et visuel, et de 8.2° pour 25°. On parle alors de *biais intersensoriels*.

L'hypothèse de la modalité pertinente (Welch et Warren, 1980; Welch, 1999) avance que la modalité sensorielle possédant la plus grande précision pour une tâche à accomplir domine la décision prise par le sujet vis-à-vis de cette tâche. Ainsi, la vision dominerait l'audition pour les tâches de localisation car la précision spatiale du système visuel est de l'ordre d'une minute d'arc (1/60e de degré) (Cavonius et Robbins, 1973) tandis que celle du système auditif est supérieure ou égale à 5° (Recanzone *et al.*, 1998) selon la nature du stimulus et son intensité.

En accord avec l'hypothèse de la modalité pertinente, le modèle du « maximum de vraisemblance » a été utilisé avec succès dans plusieurs études pour estimer la perception d'un stimulus audiovisuel disparate. Le modèle prend en compte la précision des informations spatiales visuelle et auditive en condition unimodale (Ernst et Banks, 2002; Alais et Burr, 2004). Si un stimulus visuel est perçu à une position m_V avec une variance σ_V^2, et un stimulus sonore est perçu à une position m_A avec une variance σ_A^2, alors le modèle prédit qu'un stimulus audiovisuel sera perçu à une position m_{AV} telle que :

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

Ainsi, l'influence de chaque modalité sur la localisation du stimulus audiovisuel est pondérée en fonction de sa précision en condition unimodale. ftâche Si la précision visuelle décroît,

alors la pondération associée à m_V décroît et la pondération associée à m_A augmente. Ainsi, le stimulus audiovisuel est perçu plus proche du stimulus sonore. Le modèle suggère donc que le stimulus sonore peut dans certains cas capturer le stimulus visuel si ce dernier est perçu avec moins de précision. Cet effet ventriloque « inverse » a pu être observé par Alais et Burr (2004) : les sujets devaient localiser des « taches » de lumière associées à des clics sonores. Les taches de lumières pouvaient être plus ou moins floutées. Avec les taches nettes, la localisation visuelle était précise et la source visuelle « capturait » la source sonore. Cependant, lorsque les taches de lumière était sévèrement floutées, la localisation visuelle était mauvaise et la source sonore « capturait » la source visuelle. Pour des taches moins floutées, aucun des deux sens ne dominait l'autre et le stimulus audiovisuel était perçu au milieu des stimuli sonore et visuel.

Inversement, si la précision auditive décroît, alors le système visuel dominera d'autant plus la perception spatiale de la source sonore.

3.1.2 Effet ventriloque

Dans une tâche de discrimination, les sujets indiquent si le stimulus visuel et le stimulus sonore « fusionnent » entre eux (Jack et Thurlow, 1973; Radeau et Bertelson, 1977; Thurlow et Jack, 1973) ou, dans une formulation légèrement différente, si ils perçoivent que les stimuli sonore et visuel proviennent ou non de la même direction (Choe *et al.*, 1975; André *et al.*, 2014; Werner *et al.*, 2013; Wallace *et al.*, 2004). Les études résument en général les performances des sujets en indiquant le « seuil à 50% » (qui correspond à l'écart angulaire entre stimuli sonore et visuel associés pour lequel les sujets trouvent que les stimuli fusionnent - ou perçoivent que les stimuli proviennent du même endroit - une fois sur deux). Dans certaines études, les sujets doivent plutôt utiliser l'échelle de dégradation ITU-R à 5 notes (ITU-R BS.1284-1, 2003) pour quantifier à quel point les disparités sont perceptibles et gênantes (Komiyama, 1989; Bruijn et Boone, 2002; Melchior *et al.*, 2003, 2006; Mannerheim, 2011). Comme on peut le voir sur le Tableau 3.1, les notes attribuées par les sujets sont associées à des impressions subjectives bien spécifiques. Les expérimentateurs indiquent en général le seuil pour un écart angulaire « perceptible, mais non gênant » (Mannerheim, 2011), ou alors celui pour un écart angulaire « légèrement gênant » (Melchior *et al.*, 2003, 2006). Komiyama (1989) définit également la « limite d'erreur angulaire acceptable » comme étant la frontière entre « perceptible, mais non gênant » et « légèrement gênant ».

L'effet ventriloque (observé dans des tâches de discrimination) et les biais intersensoriels (observés dans des tâches de localisation) sont deux phénomènes différents. La relation entre eux a été étudiée par Bertelson et Radeau (1981) et Wallace *et al.* (2004), qui ont mené en parallèle des tâches de localisation et de discrimination en utilisant de courtes salves de sons purs (300Hz) ou de bruits à large bande associées à des LEDs. Les deux études ont montré

Qualité	Note
Imperceptible	5.0
Perceptible, mais non gênant	4.0
Légèrement gênant	3.0
Gênant	2.0
Très gênant	1.0

Tableau 3.1 – Échelle de dégradation ITU-R à 5 notes (ITU-R BS.1284-1, 2003).

que des biais de localisation auditive vers le stimulus visuel pouvaient être observés même quand le sujet ne percevait pas de fusion et que les biais était plus importants lorsque le sujet percevait une fusion.

3.1.3 L'effet ventriloque en azimut

Les nombreuses études sur l'effet ventriloque en azimut sont toutes arrivées à la conclusion que l'effet décroit lorsque l'écart angulaire entre les stimuli sonore et visuel augmente. Cependant, les seuils rapportés sont très variables d'une étude à l'autre : de 3° pour Lewald *et al.* (2001) jusqu'à 20° avec Komiyama (1989). Ces différences peuvent être expliquées par plusieurs facteurs :

— l'expérience du sujet ;
— la disparité temporelle entre le son et l'image ;
— le réalisme de la combinaison son-image ;
— l'attention du sujet.
— l'utilisation de sources réelles ou virtuelles ;

L'expérience du sujet

Dans l'étude de Komiyama (1989), les sujets devaient utiliser l'échelle de dégradation ITU-R à 5 notes pour quantifier à quel point les disparités entre la bouche d'un personnage (diffusée sur une télévision) et sa voix leur paraissait perceptible et gênant. Avec des sujets experts, la « limite d'erreur angulaire acceptable » était de 11°, alors qu'elle atteignait 20° avec des sujets naïfs. Une comparaison des études précédentes montre également que les seuils obtenus avec des sujets entraînés sont en général bas ($5-7°$ pour Melchior *et al.* (2003), $4-8°$ pour Melchior *et al.* (2006), $6-8°$ pour Mannerheim (2011)) alors qu'ils dépassent toujours les 10° avec des sujets naïfs (20° pour Komiyama (1989), 18.3° pour André *et al.* (2014), $> 10°$ pour Wallace *et al.* (2004)).

Disparité temporelle

Retarder ou avancer le son par rapport à l'image peut également affecter l'efficacité de l'effet ventriloque. En utilisant un bruit à large bande de 50 ms suivi par l'allumage pendant 50 ms également d'une LED, Wallace a obtenu des seuils à 50% supérieurs à 15° quand le délai entre les deux stimuli étaient de 200 ms. Cependant, le seuil était réduit à 10° quand le délai atteignait 800 ms (Wallace *et al.*, 2004).

Réalisme de la combinaison son-image

L'efficacité de l'effet ventriloque dépend également de facteurs de plus haut niveau : par exemple, l'effet marchera mieux si le sujet présume que les stimuli sonore et visuel vont « bien ensemble » (Vatakis et Spence, 2007). En effet, plusieurs études ont montré que si l'association entre un stimulus sonore et un stimulus visuel est perçue comme étant plus « réaliste », les sujets considéreront plus facilement que les stimuli forment un seul et même stimulus audiovisuel (Jackson, 1953; Welch et Warren, 1980) et qu'ils ont donc une origine spatiale commune.

Par exemple, Jackson (1953) a conduit une expérience dans laquelle le stimulus visuel était de la vapeur s'échappant d'une bouilloire, associée à un son de sifflement. Il a également conduit une seconde expérience dans laquelle l'association image-son était beaucoup plus abstraite, puisqu'il s'agissait de sons de cloches associés à des allumages/extinctions de lumière :

— quand l'écart angulaire entre son et image était de 30°, le son de sifflement était encore perçu sur la bouilloire dans 97% des cas ;
— quand l'écart angulaire entre son et image était de 22.5°, le son de cloche n'était perçu sur la lumière que dans 43% des cas.

Dans une des expériences de Thurlow et Jack (1973), une télévision était placée devant le sujet, et une enceinte était cachée 20° sur la gauche de la télévision. Le stimulus visuel pouvait être soit un personnage en train de lire des textes, soit une croix au milieu d'un cercle, dessinée sur une feuille de papier blanc (le papier était collé sur la télévision, avec la croix placée au même endroit que la bouche du personnage). Dans les deux conditions, le stimulus sonore était la voix du personnage en train de lire des textes. Les sujets devaient fixer la bouche du personnage (ou la croix à l'intérieur du cercle), enclencher un chronomètre lorsqu'ils entendaient le son de la voix provenir de la même direction que le stimulus visuel, ou l'éteindre lorsqu'ils entendaient le son de la voix provenir d'une autre direction (le stimulus durait 5 minutes). La durée moyenne durant laquelle la voix était perçue dans la même direction que le stimulus visuel était :

— de 3 minutes et 22 secondes avec le personnage ;
— de seulement 51 secondes avec la croix.

Dans une étude de Warren *et al.* (1981), le stimulus visuel était soit le visage d'un personnage sur un écran, soit un petit morceau de ruban adhésif (1 × 2 cm) placé sur l'écran au même endroit que la bouche du personnage. Le stimulus sonore était la voix du personnage. Les seuils à 50% obtenus étaient plus larges avec le visage du personnage (4.6°) qu'avec le morceau de ruban adhésif (3.2°).

Ainsi, l'effet ventriloque sera plus efficace avec des séquences réalistes qu'avec des séquences abstraites.

Attention du sujet

Plusieurs études ayant mené des tâches de discrimination et de localisation en parallèle suggèrent que focaliser l'attention sur la position de la source sonore rend les sujets bien plus discriminants et atténue donc l'effet ventriloque substantiellement. En effet, de nombreux sujets ont rapporté avoir remarqué des incohérences spatiales entre son et image uniquement lorsqu'ils devaient localiser la source sonore (Radeau, 1974; Radeau et Bertelson, 1976), ou lorsque les expérimentateurs leur disaient qu'ils pourraient à tout moment pendant le test être interrogés sur la position de la source sonore (Bertelson et Radeau, 1981). Bertelson et Radeau ont formulé l'hypothèse que l'être humain, à moins d'être interrogé dessus, avait tendance à ignorer l'origine spatiale des informations auditives au profit des informations visuelles. Thurlow et Jack (1973) ont également recommandé de privilégier les tâches de discrimination aux tâches de localisation, car elles sont plus valides écologiquement et empêchent le sujet d'être exagérément discriminant.

Cependant, lorsque Komiyama (1989) a conduit sa tâche de discrimination, plusieurs sujets ont rapporté qu'ils n'auraient probablement pas remarqué les disparités audiovisuelles si on ne leur avait pas posé la question. Ces observations suggèrent que même une tâche de discrimination est peut être trop discriminante par rapport à la « vraie vie ».

Sources réelles vs. sources virtuelles

La plupart des études citées ci-dessus utilisent des sources réelles, sauf les études de Melchior *et al.* (2003; 2006), d'André *et al.* (2014) et de Mannerheim (2011) qui utilisent un système WFS. Il est cependant difficile d'évaluer l'impact de ce système sur l'effet ventriloque par rapport à des sources réelles, car les différences de résultats entre études peuvent être plutôt dues aux différences de stimuli, de protocole, de formulation de la question posée, etc.

Seeber et Fastl (2004) ont étudié l'effet ventriloque en comparant des sources réelles avec des sources virtuelles reproduites en binaural (HRTF individualisées) mais ils n'ont pas observé de différences significatives entre les deux modes de reproduction.

3.1.4 L'effet ventriloque en élévation

L'effet ventriloque a été très largement étudié dans le plan horizontal, et dans une moindre mesure en distance (Gardner, 1968; Mershon *et al.*, 1980; Agganis *et al.*, 2010; Zahorik, 2003; Hládek *et al.*, 2013; Bowen *et al.*, 2011). Par contre, très peu d'études se sont intéressées à l'élévation (Thurlow et Jack, 1973; Werner *et al.*, 2013).

Il a été montré précédemment que les biais intersensoriels de la localisation auditive vers le stimulus visuel étaient de plus en plus larges lorsque la précision de localisation auditive diminuait. Comme les performances de localisation dans le plan vertical sont bien moins bonnes que dans le plan horizontal pour des sources frontales, les biais intersensoriels devraient donc être plus importants en élévation qu'en azimut. En effet, lorsque Makous et Middlebrooks (1990) ont demandé à des sujets de localiser des bruits à large bande dont la position pouvait varier dans les deux dimensions, la variabilité intra-sujet des réponses (c'est-à-dire l'écart-type des réponses de part et d'autre de la réponse moyenne) était 2.5 fois plus importante en élévation qu'en azimut.

Si des biais plus importants sont obtenus en élévation qu'en azimut, alors, selon les études de Bertelson et Radeau (1981) et de Wallace *et al.* (2004) comparant biais intersensoriels et effet ventriloque, l'effet ventriloque devait être plus efficace en élévation qu'en azimut.

Dans une des expériences de Thurlow et Jack (1973), une télévision était placée sur le sol, droit devant le sujet, de telle sorte que la bouche du personnage à l'écran était environ 40° en-dessous du niveau des oreilles. Une enceinte était également cachée 55° au-dessus de la télévision. La durée moyenne durant laquelle le son de la voix était perçu sur la bouche du personnage était d'environ 4 minutes sur 5. La durée était réduite à 1 minute et 42 secondes lorsque le personnage à l'écran était remplacé par une croix dessinée sur une feuille de papier.

Dans une autre expérience, l'enceinte cachée se trouvait derrière les sujets. L'écart angulaire dans le plan médian entre la télévision et l'enceinte cachée était de 195°. La durée moyenne durant laquelle la voix était perçue dans la même direction que la bouche du personnage à l'écran était d'environ 3 minutes et 20 secondes, soit tout de même 2/3 de la durée totale du stimulus. La durée moyenne était réduite à 74 secondes avec la croix dessinée sur une feuille de papier.

D'un autre côté, lorsque Thurlow a mené une expérience avec une enceinte cachée à 60° sur la droite de la télévision (aucune différence d'élévation), la durée moyenne durant laquelle la voix était perçue dans la même direction que la bouche du personnage à l'écran était de seulement 52 secondes, ce qui veut dire que la durée durant laquelle l'effet ventriloque fonctionnait était 4 fois plus courtes que lorsque l'enceinte se situait derrière le sujet, quand bien même l'écart angulaire était presque trois fois plus petit.

Ainsi, les expériences de Thurlow confortent l'hypothèse que l'effet ventriloque est bien plus efficace dans le plan vertical que dans le plan horizontal, et que l'effet ventriloque dépend

fortement du réalisme de la combinaison son-image.

Dans une étude de Werner *et al.* (2013), le stimulus sonore était soit une salve de bruit blanc de 30 ms, soit un enregistrement de saxophone de 6 secondes , qui pouvait être diffusé sur 2 haut-parleurs virtuels reproduits au casque en binaural. Un des haut-parleurs se trouvait droit devant le sujet (azimut 0°, élévation 0°) tandis que le second haut-parleur se trouvait à 25° au-dessus du niveau des oreilles. Les stimuli visuels étaient des allumages/extinctions de LEDs blanches, disposées à 0° d'azimut le long d'un arc de cercle centré sur le sujet, de 10° en-dessous du niveau des oreilles jusqu'à 35° au-dessus du niveau des oreilles. Pour chaque présentation, le sujet devait indiquer si la source visuelle était en-dessous, au même niveau, ou au-dessus du stimulus visuel. Il est à noter que dans cette expérience, la position du stimulus sonore était fixe tandis que celle du stimulus visuel variait.

Le seuil à 50% a été estimé à ±8° pour l'enceinte située à élévation 0° (que le stimulus visuel soit déplacé vers le haut ou vers le bas par rapport à l'enceinte). Pour la seconde enceinte, le seuil était de 10° lorsque le stimulus visuel était déplacé vers le bas et supérieur à 10° lorsque le stimulus visuel était déplacé vers le haut.

Werner en a conclu que l'effet ventriloque avait la même amplitude en élévation et en azimut. Cependant, Werner n'a pas mené son expérience dans le plan horizontal et sa conclusion repose donc sur une comparaison avec des études précédentes, dont les conditions expérimentales étaient différentes. Comme nous avons pu le voir précédemment, l'efficacité de l'effet ventriloque dépend fortement des conditions expérimentales : il n'est donc pas possible à ce stade de savoir si les similarités de seuils observées par Werner sont véritablement dues au fait que l'effet ventriloque a la même amplitude en élévation et en azimut, ou si elles ne résultent pas tout simplement de différences de protocoles expérimentaux.

Dans l'étude de Werner, de nombreux facteurs ont pu favoriser l'obtention de seuils à 50% particulièrement bas :

- les sujets étaient expérimentés et entraînés pour la tâche à accomplir. Or, il a été montré que les seuils obtenus avec des experts pouvaient être jusqu'à deux fois plus petits que ceux obtenus avec des sujets naïfs (Komiyama, 1989) ;
- le son et l'image étaient parfaitement synchrones ;
- la combinaison son-image n'était pas très réaliste puisqu'il s'agissait de salves de bruit blanc associées à des allumages/extinctions de LEDs blanches. Un enregistrement de saxophone a également été utilisé, ce qui peut paraître plus réaliste. Cependant, l'enregistrement restait associé à des allumages/extinctions de LEDs blanches, ce qui n'a probablement pas rendu la combinaison son-image plus crédible ;
- le stimulus sonore était fixe (soit à 0° d'élévation, soit à 25° d'élévation). Les sujets se sont donc probablement faits une idée de plus en plus précise au cours du test de la position des enceintes, ce qui les a rendu plus discriminants quand ils devaient comparer avec la position des stimuli visuels.

Si l'expérience de Werner avait été conduite en azimut dans les mêmes conditions, les seuils obtenus auraient sûrement été encore plus bas. Il est donc possible que les résultats de Werner aient sous-estimé la force de l'effet ventriloque en élévation par rapport à celle en azimut. Nous tenterons de vérifier ce point dans l'expérience IV (chapitre 5.1).

3.1.5 L'effet ventriloque en profondeur

Même si le phénomène a été beaucoup moins exploré qu'en azimut, l'effet ventriloque en profondeur a tout de même fait l'objet d'un nombre important d'études (Gardner (1968); Mershon *et al.* (1980); Zahorik (2003); Agganis *et al.* (2010); Zahorik (2001); Calcagno *et al.* (2012); Bowen *et al.* (2011); Turner *et al.* (2011); Côté *et al.* (2012); Corrigan *et al.* (2013), voir André (2013) pour une synthèse détaillée des études).

Gardner (1968) est le premier à avoir mis en évidence un phénomène d'attraction du son par l'image dans la profondeur : dans une chambre anéchoïque, des sujets faisaient face à 5 haut-parleurs placés les uns derrière les autres à hauteur des yeux, si bien que les sujets ne pouvaient voir que le haut-parleur le plus proche. Lorsque une source était diffusée sur le haut-parleur le plus éloigné (à environ 9 mètres), tous les sujets sans exception localisaient le son sur le haut-parleur le plus proche.

Mershon *et al.* (1980) ont reconduit une telle expérience en chambre anéchoïque, mais aussi en chambre réverbérante, et ont obtenu des résultats similaires. Ils ont également remarqué que l'effet ventriloque en profondeur était asymétrique et fonctionnait moins bien lorsque la source visuelle était plus loin que la source sonore. Cette asymétrie a été confirmée par Zahorik (2003).

Plus récemment, Agganis *et al.* (2010) ont mesuré des biais intersensoriels pour des disparités à la fois en azimut et en distance, et ont obtenu des biais plus importants dans la profondeur qu'en azimut.

3.2 Influence de la stéréoscopie sur la perception du son : Etat de l'Art

Peu d'études ont été faites sur la perception du son lié à l'image en 3D-s. Dans ce chapitre, nous comparons certains témoignages d'ingénieurs du son ayant mixé pour la 3D-s. Nous présentons également diverses études (André *et al.*, 2012; Moulin, 2015; Kruszielski *et al.*, 2012; Iljazovic *et al.*, 2012; Kamekawa *et al.*, 2011) ayant cherché à déterminer si certains systèmes de reproduction sonore semblaient plus appropriés que d'autres pour l'image en 3D-s. Alors que les études d'André et de Moulin utilisaient uniquement des contenus 3D-s, les études de Kruszielski, Illjazovic et Kamekawa ont également diffusé les séquences de leur test en 2D pour évaluer l'influence de la stéréoscopie sur les préférences des sujets.

3.2.1 À Hollywood, des opinions contradictoires...

Certains ingénieurs du son affirment que les versions 2D et 3D-s d'un film doivent être mixées différemment (Krohn, 2009). Michael Semanick, par exemple, affirme avoir ajouté bien plus d'effets, de musique, d'ambiance et de réverbération dans les enceintes « surround » et avoir plus latéralisé les dialogues pour la version 3D-s de *Alice aux pays des merveilles* de Tim Burton (Gambier, 2010). Paul Martin Smith, monteur du film de Eric Brevig *Voyage au Centre de la Terre*, défend également l'idée qu'un mixeur devrait toujours travailler avec l'image projetée en 3D-s, car cela influence la spatialisation des sources sonores (Krohn, 2009).

En revanche, les ingénieurs du son de *Hugo Cabret* (film de Martin Scorsese) estiment que l'image, avec la 3D-s, n'a fait que rattraper son retard sur le son, qui était déjà en 3D depuis des décennies avec le 5.1. Le son ne devrait donc pas être trop affecté par le « phénomène 3D » (Coleman, M., b). Quant aux ingénieurs du son d'*Avatar* (film de James Cameron), ils rapportent avoir mixé le film avec l'image projetée en 2D, et n'avoir pratiquement pas eu à faire de modifications lorsqu'ils ont vérifié leur mixage en 3D-s (Coleman, M., a).

3.2.2 Comparaisons de systèmes de reproduction sonore avec images projetées en 3D-s

André *et al.* (2012) : Influence de la cohérence audiovisuelle spatiale sur la sensation de présence

Dans cette étude, André a posé l'hypothèse qu'une reproduction du son spatialement cohérente avec l'image devrait accroitre la sensation d'être « dans le film » de la même manière que la stéréoscopie accroit la sensation de présence des spectateurs, comme l'ont montré Ijsselsteijn *et al.* (2001). Il a donc diffusé à des sujets un extrait d'un film d'animation 3D-s avec trois bandes-son différentes :

— le mixage original stéréophonique (cohérence spatiale faible) ;

— un mixage en WFS (cohérence spatiale forte en azimut. Par contre, la cohérence n'était pas forcément assurée en élévation, le système WFS étant à hauteur fixe.) ;
— un mixage hybride : il s'agissait d'un mixage stéréophonique, sauf que les objets sonores était latéralisés (à l'aide d'un « panning » d'amplitude) et plus ou moins atténués (facteur d'atténuation r^2) en fonction de l'azimut et de la profondeur de leur objet visuel correspondant.

Les résultats ont montré que la configuration sonore n'avait globalement pas eu d'impact significatif sur la sensation de présence. Cependant, André a identifié un groupe de sujets (12 sujets sur les 33) ayant rapporté d'importantes sensations de « présence » pendant le test, toute configuration confondue. En limitant les analyses à ce groupe, il a observé une influence significative de la configuration sonore, la WFS ayant procuré une moins grande sensation de présence que les deux configurations stéréophoniques.

Moulin (2015) : Influence du système de reproduction et de la cohérence audio-visuelle en distance sur l'expérience audiovisuelle

Dans un premier temps, Moulin a souhaité étudier l'influence du système de reproduction sur la qualité d'expérience audiovisuelle. Il a donc présenté à 30 sujets 15 séquences audiovisuelles extraites d'un documentaire en 3D-s. Les séquences avaient été mixées en 5.1 et étaient diffusées avec plusieurs systèmes de reproduction sonore différents :

— un système 5.1 standard ;
— un casque proposant un *down-mix* 2.0, c'est-à-dire une réduction à 2 canaux du mixage 5.1 original ;
— une barre sonore, capable de restituer le signal 5.1 grâce aux techniques de *beam-forming*. Cette technique permet de recréer des haut-parleurs virtuels placés à différentes positions, en exploitant notamment les réflexions des ondes acoustiques sur les parois de la salle de diffusion.

Les sujets devaient évaluer :

— le degré de profondeur visuelle ;
— le confort de visualisation ;
— le degré de spatialisation sonore ;
— le confort d'écoute ;
— le degré d'immersion ;
— le degré de cohérence entre le son et l'image.

Les résultats montrent que :

— le système de restitution sonore n'a pas eu d'effet sur le degré de profondeur visuelle perçu ni sur le confort de visualisation ;
— la diffusion au casque a procuré une plus grande sensation de spatialisation sonore que

le système multicanal 5.1, et la barre sonore a été le système de diffusion qui offrait le moins d'effets de spatialisation ;

— le système de restitution sonore n'a pas eu d'effet sur le confort d'écoute, qui a été plutôt bien noté (notes comprises entre 3/4 et 3.5/4). Selon Moulin, ce résultat suggère que les participants n'ont pas été gênés par le port simultané du casque audio et des lunettes 3D ;
— le degré de cohérence entre le son et l'image a été légèrement plus élevé au casque qu'avec la barre sonore. Quant au système multicanal 5.1, il n'était significativement différent ni du casque ni de la barre sonore ;
— le système de restitution sonore n'a pas eu d'effet sur le degré d'immersion ;
— Moulin estime que les critères d'immersion et de cohérence entre le son et l'image présentent une corrélation élevée (0,67).

Ce dernier point a amené Moulin à suggérer que l'amélioration de l'immersion du spectateur devait passer par la recherche d'un système de reproduction sonore procurant un degré de cohérence audiovisuelle plus important. Il ne peut pas s'agir de cohérence sémantique ou temporelle, puisque le son et l'image étaient synchrones et avaient été captés en même temps pour toutes les séquences. Moulin a donc posé l'hypothèse qu'il s'agissait de la cohérence audiovisuelle spatiale et a décidé de conduire une nouvelle expérience (il est à noter que cette hypothèse ressemble fortement à celle d'André *et al.* (2012)).

Dans cette nouvelle expérience, Moulin a cherché à savoir si une restitution physiquement réaliste de la distance d'objets sonores pouvait améliorer la qualité d'expérience des spectateurs (voir chap. 1.2.4 pour une définition de « restitution physique », qui s'oppose à la « restitution psychoacoustique » proposée par les systèmes multicanaux classiques). 9 séquences audiovisuelles ont été tournées puis présentées à 24 sujets, avec plusieurs mixages différents :

— un mixage « sans distance », c'est-à-dire sans restitution physique de la distance sonore. Les objets sonores étaient donc « classiquement » reproduits sur des enceintes cachées derrière l'écran ;
— un mixage « distance réaliste », où les sources sonores, grâce à la WFS, étaient virtuellement placées à la même distance que leur correspondant visuel ;
— un mixage « distance augmentée », où la distance des sources sonores virtuelles était volontairement exagérée par rapport à la distance de leur correspondant visuel.

Il est à noter que Moulin ne modifiait d'un mixage à l'autre ni la balance spectrale (équalisations) ni la réverbération, et n'agissait que sur le niveau sonore des sources et sur les courbures des fronts d'onde. De plus, cette étude se focalisait uniquement sur la cohérence en profondeur. Qu'importe le mixage, une source sonore était toujours diffusée au même azimut que la position à l'écran de son correspondant visuel. Par contre, la cohérence verticale n'était pas forcément assurée, le système WFS étant à hauteur fixe.

Les sujets devaient évaluer sur une échelle de 0 à 10 :

— 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 audiovisuelle : avez-vous l'impression d'être présent, immergé dans la scène ? La scène vous semble-t-elle réelle (cohérence entre son et image, etc.) ?

Les résultats ont montré que :

— Les participants ont détecté les différences entre les trois mixages sonores en termes de profondeur sonore restituée, et les notes de profondeur sonore ont été généralement supérieures pour les mixages avec restitution de la distance que pour le mixage « sans distance » ;

— la qualité sonore perçue a été globalement comparable pour les trois mixages audio ;

— Pour quelques séquences, la profondeur visuelle perçue a été légèrement plus élevée avec le mixage « distance augmentée » ;

— Le mixage sonore n'a pas eu d'effet sur la gêne visuelle ;

— la cohérence audiovisuelle en profondeur (mixage « distance réaliste ») n'a amélioré la sensation d'immersion des sujets que pour 2 séquences sur 9 par rapport à une diffusion monophonique «classique» sur enceintes cachées derrière l'écran (mixage « sans distance », voir Fig. 3.1).

3.2.3 Comparaisons de systèmes de reproduction sonore avec images à la fois projetées en 3D-s et en 2D

Les études de Moulin et André ont permis de comparer les performances de différents systèmes de reproduction sonore lorsqu'ils sont combinés à des images en 3D-s. Bien que leurs recherches aient été orientées par les spécificités propres à la 3D-s (restitution visuelle de la profondeur qui pousse à envisager également une restitution physiquement réaliste de la profondeur sonore, etc.), rien ne permet de savoir si leurs résultats sont propres à la stéréoscopie et si des tendances totalement différentes auraient été observées avec des images en 2D.

Les études de Kruszielski, Iljazovic et Kamekawa ont la particularité d'avoir diffusé les séquences de leur test à la fois dans leur version 3D-s et 2D, ce qui a permis d'évaluer l'influence de la stéréoscopie sur les jugements des sujets.

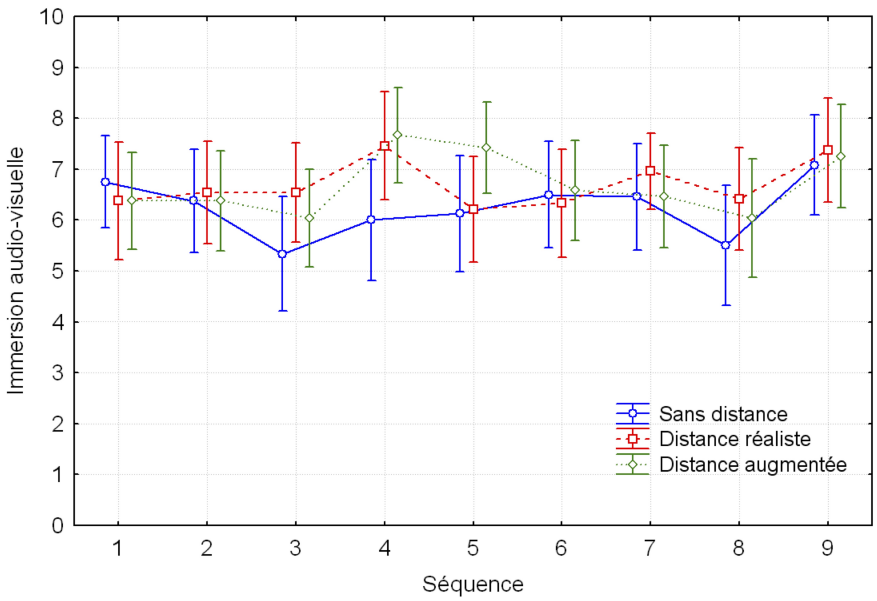


FIGURE 3.1 – Notes d'immersion audiovisuelle et intervalles de confiance à 95% associés pour les trois mixages sonores : « sans distance » (en bleu), « distance réaliste » (en rouge) et « distance augmentée » (en vert). D'après Moulin (2015).

Kruszielski *et al.* (2012) : Evaluation de la distance et de l'adéquation du son à l'image en fonction du placement du système d'enregistrement par rapport à la caméra pour une image 2D et 3D-s

Kruszielski a filmé un saxophoniste en plaçant la caméra 3D à 5 distances différentes du musicien (voir Fig. 3.2). Pour chaque plan, les sujets devaient juger à quel point les différentes bandes-son qui leur étaient présentées étaient adaptées (« suitable » en anglais) à l'image. Ces différentes bandes-son avaient été obtenues en plaçant également un système d'enregistrement à différentes distances du musicien. La diffusion pouvait se faire soit en stéréo, soit en 5.1. Pour chaque présentation, les sujets devaient évaluer leur sensation de distance et quantifier à quel point le son leur paraissait adapté à l'image.

FIGURE 3.2 – Images obtenues pour les 5 positions de la caméra. D'après Kruszielski *et al.* (2012).

Les résultats ont montré pour la sensation de distance que :

— le saxophoniste paraissait plus loin avec la diffusion 5.1 qu'avec la diffusion stéréophonique ;

— pour la moitié des sujets, le saxophoniste était perçu plus loin lorsque l'image était projetée en 3D-s, et leur sensation de distance était principalement déterminée par la position de la caméra ;

— pour l'autre moitié des sujets, la sensation de distance était la même en 2D et en 3D-s, et leur sensation de distance était principalement déterminée par la position du système d'enregistrement.

La sensation de distance fait donc apparaître deux groupes présentant deux stratégies différentes de notation. Pour la sensation d'adéquation du son à l'image, les résultats ont montré que :

— le système d'enregistrement proche de la position de la caméra était toujours jugé comme étant le plus adapté ;

— plus les systèmes d'enregistrement avaient été placés loin de la caméra, moins ils étaient jugés « adaptés » ;

— le mode visuel (2D vs 3D-s) n'a pas eu d'impact sur l'adéquation du son à l'image ;

— la diffusion 5.1 a été perçue comme plus adaptée que la diffusion stéréophonique ;

— les deux groupes de sujets mis en évidence pour la sensation de distance ont répondu de la même manière pour l'adéquation du son à l'image.

Ainsi, quand bien même certains sujets percevaient le saxophoniste plus loin lorsque l'image était projetée en 3D-s, l'adéquation du son à l'image était la même en 2D et en 3D-s.

Iljazovic *et al.* (2012) : Influence du rendu spatial sur la qualité du son perçue au casque pour une image 2D et 3D-s

Iljazovic a montré à 45 sujets 6 séquences extraites de films, de documentaires et de concerts filmés. Les sujets étaient divisés en trois groupes de 15 sujets :

— au premier groupe n'était diffusé que le son des séquences, sans image ;

— le deuxième groupe regardait les séquences en 2D ;

— le troisième groupe regardait les séquences en 3D-s.

Le son était diffusé au casque sous trois formats différents : 5.1 (« binauralisé » grâce à l'un des 3 algorithmes de traitement que l'étude se proposait de comparer : Cond. A, Cond. B et Cond. C.), stéréophonie ou alors monophonie. Les sujets devaient évaluer la qualité du son sur une échelle à 100 points allant de « pauvre » à « excellent ». Les résultats ont montré que :

— le son paraissait de meilleure qualité avec l'image 2D que sans image ;

— le son paraissait de moins bonne qualité avec l'image 3D-s comparé aux conditions

« avec image 2D » et « sans image ». Selon l'auteur, ce résultat pourrait être dû au fait que les attentes concernant la qualité spatiale de la reproduction sonore sont plus élevées lorsqu'une vidéo est projetée en 3D-s ;

— les sujets ont globalement préféré les mixages 5.1 aux mixages stéréophoniques traditionnels ;

— cette préférence était significativement plus marquée lorsque les séquences étaient projetées dans leur version 3D-s (voir Fig. 3.3).

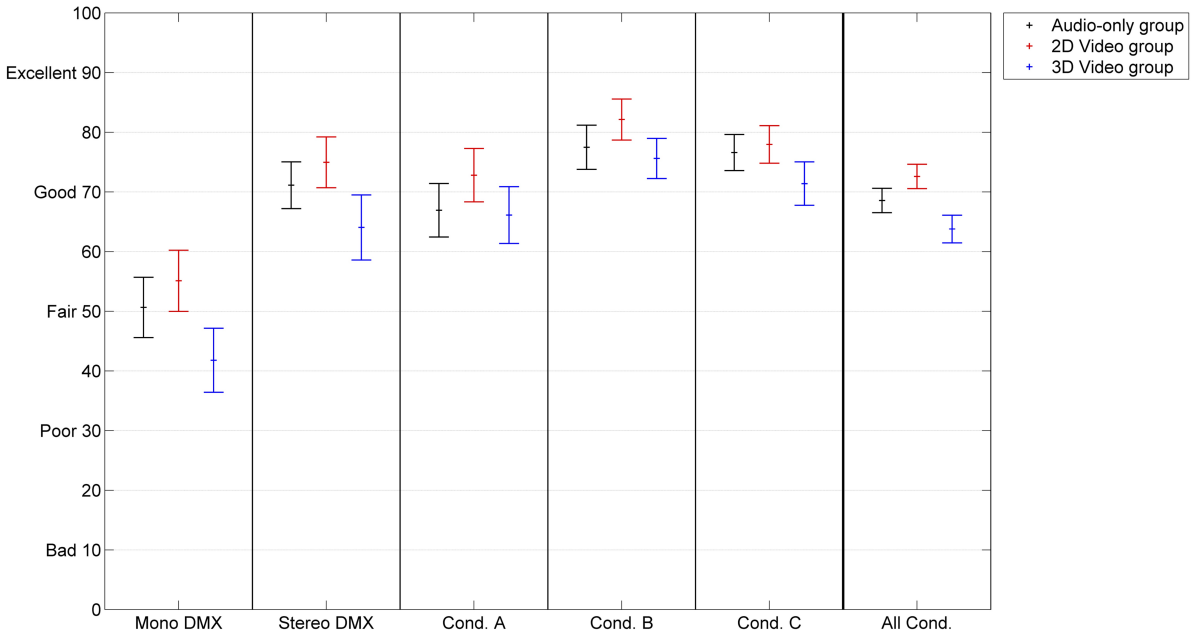


FIGURE 3.3 – Notes et intervalles de confiance à 95% pour chaque format sonore (mono, stéréo, et 5.1 « binauralisé » grâce à 3 algorithmes de traitement différents : Cond. A, Cond. B et Cond. C). D'après Iljazovic *et al.* (2012).

Kamekawa *et al.* (2011) : Influence du système de reproduction sur la sensation de profondeur et d'adéquation du son à l'image pour une image 2D et 3D-s

Kamekawa a montré à 13 sujets 2 séquences musicales avec trois systèmes de reproduction différents : stéréo, 5.0 et 7.0 (c'est-à-dire un 5.0 avec deux enceintes zénithales supplémentaires, l'une au dessus de l'enceinte gauche et l'autre au-dessus de l'enceinte droite). Les sujets devaient évaluer sur une échelle de 1 à 7 :

— la sensation de distance sonore ;

— l'adéquation du son à l'image (« sound suitability »).

Les résultats ont montré que :

— les mixages 7.0 ont donné les sensations de distance les plus élevées ;

— les mixages ont globalement donné des sensations de distance plus élevées lorsqu'ils étaient présentés avec l'image en 3D-s ;

— les mixages 5.0 et 7.0 avec image 2D étaient perçus plus lointains que le mixage stéréophonique avec image en 3D-s ;

— les mixages 7.0 avec image en 3D-s ont donné les sensations d'adéquation du son à l'image les plus élevées ;

— pour l'image en 3D-s, les mixages 7.0 et 5.0 ont été jugés plus adaptés que les mixages stéréophoniques, alors que pour l'image 2D, il n'y a pas de différence significative entre mixages stéréophoniques, 5.0 et 7.0 sur l'adéquation du son à l'image (voir Fig. 3.4).

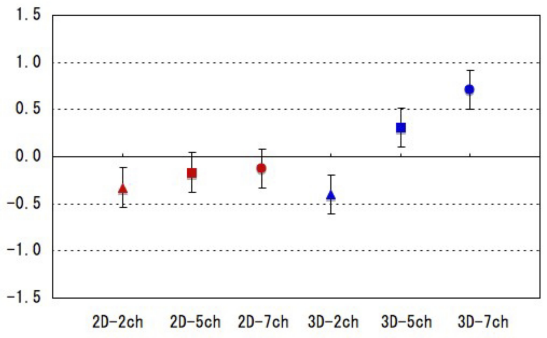


FIGURE 3.4 – Notes et intervalles de confiance à 95% pour l'adéquation du son en fonction du Mode Visuel (2D vs. 3D-s) et du système de reproduction (stéréo, 5.0, 7.0). D'après Kamekawa *et al.* (2011).

3.3 Conclusion

L'hypothèse générale que nous souhaitons vérifier dans le cadre de notre étude est :

Hypothèse globale : la stéréoscopie modifie significativement notre perception ou nos attentes sonores par rapport à une projection en 2D.

Nous avons vu que cette hypothèse donnait déjà lieu à de nombreuses contradictions :

— à Hollywood, certains ingénieurs du son affirment que les versions 2D et 3D-s d'un film doivent être mixées différemment tandis que d'autres considèrent l'influence de la stéréoscopie comme étant négligeable ;

— Iljazovic et Kamekawa ont obtenu un effet significatif de la stéréoscopie sur les jugements de leurs sujets alors que l'effet était négligeable dans l'étude de Kruszielski.

Nous avons décidé pour vérifier cette hypothèse globale de conduire 5 expériences explorant deux axes différents : un premier axe concernant les sons d'ambiance et les enceintes surround, et un second concernant la spatialisation des dialogues et effets sonores. Nous définissons alors deux sous-hypothèses :

Sous-hypothèse 1 - La stéréoscopie change nos attentes en termes de balance frontal/surround : en 3D-s, nous souhaitons entendre « plus de surround ».

— Michael Semanick affirme avoir ajouté bien plus d'effets, de musique, d'ambiance et de réverbération dans les enceintes « surround » pour la version 3D-s d'*Alice aux pays des merveilles* ;

— Dans les études d'Iljazovic et de Kamekawa, la préférence des sujets pour les mixages avec surround (5.0, 5.1, 7.0) par rapport aux mixages sans surround (mono, stéréo) était plus marquée lorsque les images étaient projetées en 3D-s que lorsqu'elles étaient projetées en 2D.

— Dans l'étude de Kruszielski, les sujets préféraient également les mixages 5.1 aux mixages 2.0, mais cette préférence n'était pas plus marquée en 3D-s qu'en 2D.

Nous tenterons de vérifier la sous-hypothèse 1 dans les expériences I, II et III (voir chap. 4).

Sous-hypothèse 2 - La stéréoscopie change nos attentes en termes de spatialisation des objets sonores : en 3D-s, une plus grande cohérence spatiale entre le son et l'image est attendue.

— Michael Semanick affirme avoir plus latéralisé les dialogues pour la version 3D-s d'*Alice aux pays des merveilles* que pour la version 2D ;

— Paul Martin Smith, monteur du film *Voyage au Centre de la Terre*, estime qu'un mixeur devrait toujours travailler avec l'image projetée en 3D-s, car cela influence la

spatialisation des sources sonores ;

— André a posé l'hypothèse qu'une reproduction du son spatialement cohérente avec l'image devrait accroitre la sensation d'être « dans le film » de la même manière que la stéréoscopie accroit la sensation de présence des spectateurs. Cependant, lorsqu'il a diffusé une séquence 3D-s avec plusieurs configurations sonores, les sujets ont trouvé que la WFS (cohérence spatiale forte) leur procurait une sensation de présence semblable ou même moins grande qu'un mixage stéréophonique classique (cohérence spatiale faible) ;

— Dans une première expérience, Moulin a observé que la sensation d'immersion était corrélée à la sensation de cohérence entre son et image et a supposé que l'amélioration de l'immersion du spectateur devait passer par la recherche d'un système de reproduction sonore procurant un degré de cohérence audiovisuelle spatiale plus important. Dans une seconde expérience, Moulin a montré que la cohérence audiovisuelle en profondeur n'améliorait la sensation d'immersion des sujets que pour 2 séquences sur 9 ;

— Dans l'étude de Kruszielski, la cohérence audiovisuelle en profondeur améliorait l'adéquation du son à l'image. Cependant, cette tendance n'était pas spécifique à la stéréoscopie, puisque l'amélioration était la même en 3D-s et en 2D.

Nous tenterons de vérifier la sous-hypothèse 2 dans les expériences IV et V (voir chap. 5).

Dans l'expérience IV, nous nous intéresserons à la cohérence audiovisuelle **en élévation**. Dans le chapitre 3.1.4, nous avons vu que certaines études suggéraient que l'effet ventriloque était particulièrement puissant dans le plan vertical : la cohérence audiovisuelle en élévation est donc peut-être inutile au cinéma.

Dans l'expérience V, nous nous intéresserons à la cohérence audiovisuelle **en azimut et en profondeur**.

Critères pour les stimuli de la présente étude

Nous constatons que les témoignages d'ingénieurs du son et les résultats d'études précédentes sont variés voire parfois contradictoires. Cette constatation suggère que l'effet de la stéréoscopie est étroitement lié au contenu des séquences utilisées dans les tests. Or, la plupart des études précédemment citées n'utilisent qu'une ou deux séquences (André *et al.*, 2012; Kamekawa *et al.*, 2011) ou alors utilisent plusieurs séquences mais qui sont fortement similaires : dans l'étude de Moulin (2015) par exemple, les 9 séquences utilisées ont été tournées dans le même décor, avec les mêmes acteurs et avec le même point de vue pour la caméra. Dans l'étude de Kruszielski *et al.* (2012), il s'agit du même saxophoniste, jouant dans la même salle, mais filmé à 5 distances différentes. Enfin, certaines des études précédemment citées utilisent intégralement (Kruszielski *et al.*, 2012; Kamekawa *et al.*, 2011) ou en partie (Iljazovic *et al.*, 2012) des séquences musicales. Or, il est probable que les attentes d'un sujet en termes d'ex-

périences audiovisuelles soient totalement différentes lorsqu'il regarde un film. Ces diverses remarques confirment le besoin pour notre étude d'utiliser **un plus grande nombre et une plus grande variété** de séquences **extraites de films**.

Selon Kunka et Kostek (2013), les productions professionnelles en 3D exploitent principalement des disparités positives pour le confort du public (voir chap. 2.2.2). Les objets visuels de nos séquences seront donc principalement situés **dans le plan de l'écran** ou **derrière l'écran**.

Deuxième partie

Contributions de la thèse

Chapitre 4

Influence de la stéréoscopie sur la perception des sons d'ambiance

Sommaire

4.1 Introduction

Nous souhaitons dans ce chapitre vérifier la sous-hypothèse 1, à savoir que la stéréoscopie change nos attentes en termes de balance frontal/surround (c'est-à-dire l'équilibre entre les enceintes frontales et les enceintes « surround »).

Michael Semanick affirme en effet avoir ajouté bien plus d'effets, de musique, d'ambiance et de réverbération dans les enceintes « surround » pour la version 3D-s d'*Alice aux pays des merveilles* de Tim Burton. En revanche, les ingénieurs du son de *Hugo Cabret* ou d'*Avatar* estiment que l'influence de la stéréoscopie est négligeable.

Nous avons également vu dans certaines études (voir chap. 3.2) que les sujets préféraient les mixages avec surround (5.1, 7.1) aux mixages sans surround (2.0), et que cette préférence était encore plus marquée lorsque les images étaient projetées en 3D-s. Cependant, d'autres études ont obtenu une influence négligeable de la stéréoscopie sur cette préférence.

Notre première série d'expériences (I, II et III) se concentre sur la perception des sons d'ambiance, puisqu'il s'agit de la catégorie de sons principalement diffusée dans les enceintes « surround ». Ces trois expériences ont fait l'objet d'une publication dans (Hendrickx *et al.*, 2014).

4.2 Expérience I : Influence de la stéréoscopie sur le mixage de sons d'ambiance

Dans un premier temps, un panel de sujets devait mixer les sons d'ambiance de plusieurs courtes séquences. Le but de l'expérience était de vérifier si leur stratégie de mixage était différente selon que l'image était projetée en 2D ou en 3D-s.

4.2.1 Matériel et méthode

.

Lieu de l'expérience

Le test s'est déroulé dans l'auditorium de mixage du département « Image et Son » de l'Université de Brest. La salle est spécialement conçue pour la post-production cinématographique, avec 5 enceintes actives professionnelles (PSI Audio A25-M) en configuration 5.0 :

— 3 enceintes frontales : enceinte frontale gauche, enceinte centrale, enceinte frontale droite ;

— 2 enceintes « surround » : enceinte surround gauche, enceinte surround droite.

Les enceintes étaient alimentées par une interface RME Fireface 800 connectée à un ordinateur Apple MacBook Pro, et leur gain avait été calibré de manière à ce qu'un bruit rose

diffusé à - 20 dBfs RMS produise à la position d'écoute de référence un niveau de pression sonore égal à 85 dBC par enceinte.

L'image (25 i/s) était projetée par un projecteur numérique Epson EH-TW6000, synchronisé avec des lunettes 3D-s actives Epson ELPGS01. Le champ visuel des images projetées était de 45° (angle idéal proposé par Dolby (1994)).

Séquences

11 séquences 3D-s, ainsi que leur version 2D, ont été utilisées pour le test. Nous avons filmé ces séquences spécialement pour le test (à l'aide d'une caméra Panasonic AG-3DP1), de manière à couvrir une large gamme de dynamiques (plans statiques, travelling avant, caméra-épaule, etc.), de valeurs (plans serrés, plans moyens, plans larges, etc.), de décors (mer, ville, foule, intérieur, forêt, etc.) et de situations (mer calme, dialogues, séquence musicale, etc.).

Nous avons également décidé d'utiliser différentes techniques d'enregistrement et de mixage qui sont fréquemment utilisées dans des productions professionnelles :

— 3 séquences ont été enregistrées avec un système double-M/S (Wittek *et al.*, 2006) ;

— 4 séquences ont été mixées en utilisant deux ambiances stéréophoniques décorrélées (une reproduite sur les enceintes frontale gauche et frontale droite, une reproduite sur les enceintes surround gauche et surround droite). Il s'agit d'une pratique très courante dans le monde du cinéma. Ces ambiances ont été enregistrées avec un couple ORTF (deux microphones cardioïdes formant un angle de 110° avec des capsules espacés de 17 cm) (Hugonnet et Jouhaneau, 1987) ;

— 2 séquences ont été enregistrées avec un système Double-ORTF (Czyzewski *et al.*, 2002) ;

— 2 séquences ont été enregistrées avec un système Fukada Tree (Hiekkanen *et al.*, 2007) (voir Fig. 4.1) ;

Ces systèmes d'enregistrement sont présentés plus en détail dans l'annexe B. La fréquence d'échantillonnage était de 48 kHz et la quantification était effectuée sur 24 bits.

Les microphones utilisés étaient constitués de corps Shoeps CMC6, avec des capsules MK4 pour les directivités cardioïdes et MK8 pour les directivités « figure en 8 » (bidirectionnels). Ils étaient reliés à un enregistreur Sonosax SX-R4. Un microphone canon Neumann KMR 81 a également été utilisé pour les enregistrements monophoniques des dialogues et des bruits de pas lors des séquences 4, 5 et 6.

Les séquences sont plus largement détaillées dans l'annexe C.

Balance sonore

Nous appelons balance « frontal/surround » le rapport de niveau sonore entre la paire d'enceintes frontales (enceintes frontale gauche et frontale droite) et la paire d'enceintes sur-

FIGURE 4.1 – Photo du tournage des séquences 1 et 2 de l'expérience I, avec la caméra 3D-s, le rail de travelling ainsi que le système Fukada Tree (5 microphones cardioïdes).

round (l'enceinte centrale C est mise de côté dans cette étude et ne servira qu'à l'éventuelle diffusion de dialogues ou autre effets sonores monophoniques).

Pour chaque séquence, deux balances frontal/surround ont été fixées par les expérimentateurs : une pour la version 3D-s, et une autre pour la version 2D. La moyenne des deux balances a ensuite été calculée afin de définir une balance frontal/surround « nominale » (fixée à 0 dB) qui servirait de référence pour les analyses à venir.

Un niveau global d'écoute a également fixé par les expérimentateurs pour chaque séquence (une pratique courante dans les tests subjectifs (IEC 60268-13, 1998)).

Protocole

Les 11 séquences 3D-s, ainsi que leur version 2D, ont été présentées aléatoirement à tous les sujets et dans un ordre différent pour chaque sujet. Les sujets étaient priés de garder leurs lunettes 3D-s pendant la totalité de l'expérience, même pour les séquences 2D (durant lesquelles la même image était envoyée à l'œil gauche et à l'œil droit), afin d'éviter une éventuelle influence de la perte de luminosité (qui peut aller de 40 à 70% avec des lunettes actives). Le sujet devait passer le test 2 fois, avec une pause de 15 minutes entre les deux sessions. La durée moyenne d'une session était de 40 minutes. Aucun sujet n'a rapporté avoir subi de fatigue visuelle ou d'inconfort pendant le test.

Les sujets devaient répondre à la question « Quelle balance avant/arrière souhaiteriez-vous pour cette séquence ? » en réglant eux-même la balance, à l'aide d'un bouton rotatif issu

d'une surface de contrôle Digidesign Command-8 (une interface MIDI qui permettait 128 pas de réglages différents pour la balance). Le bouton était « à rotation infinie » et les gradations l'entourant avaient été effacées, afin qu'aucun retour visuel ou tactile ne puisse influencer le choix du sujet. En tournant le bouton :

— dans le sens des aiguilles d'une montre, le sujet augmentait le niveau des enceintes frontale gauche et frontale droite tout en baissant le niveau des enceintes surround gauche et surround droite, jusqu'à ce qu'il n'y ait plus que du son à l'avant ;

— dans le sens inverse des aiguilles d'une montre, le sujet diminuait le niveau des enceintes frontale gauche et frontale droite tout en augmentant le niveau des enceintes surround gauche et surround droite, jusqu'à ce qu'il n'y ait plus que du son dans les « surround » ;

Une simple loi de panoramique « sinus-cosinus » fut choisie pour l'évolution de l'intensité, car elle produisait des variations d'intensité plus naturelles et plus proches de l'expérience de mixage quotidienne des sujets.

Soit n la valeur MIDI (comprise entre 0 et 127) fixée par le sujet, et G_F et G_R les gains d'amplification (exprimés en dB) appliqués respectivement aux enceintes frontales et aux enceintes surround. Alors :

$$G_F = 20 \times \log_{10}[\sin(\frac{\pi}{2} \times \frac{n}{127})] + 3 \text{ dB},$$

$$G_R = 20 \times \log_{10}[\cos(\frac{\pi}{2} \times \frac{n}{127})] + 3 \text{ dB},$$

$$\Delta G = G_F - G_R = 20 \times \log_{10}[\tan(\frac{\pi}{2} \times \frac{n}{127})].$$

Par exemple, si le sujet réglait le bouton sur sa position médiane ($n = 64$), alors $G_F = G_R \approx 0$ dB. Le sujet ne modifiait ni le niveau des enceintes frontales, ni le niveau des enceintes surround : il choisissait donc la balance telle qu'elle avait été fixée initialement par les expérimentateurs. Par contre, si le sujet tournait le bouton jusqu'à sa position maximale ($n = 127$), alors $G_F = +3$ dB and $G_R = -\infty$, et la totalité de l'énergie sonore était alors diffusée sur les enceintes frontales.

Les sujets ne pouvaient contrôler que la balance frontal/surround et ne pouvaient pas la faire évoluer pendant la séquence (il devait choisir un seul et même réglage de balance pour chaque présentation). Ils ne pouvaient pas non plus modifier le niveau de l'enceinte centrale, sur laquelle étaient diffusés les dialogues et les bruits de pas des séquences 4, 5 et 6.

Chaque séquence durait environ 30 secondes et était automatiquement mise en boucle. Le sujet modifiait les gains tout en regardant la séquence. Lorsque le sujet était satisfait de sa balance, il devait appuyer sur un bouton pour accéder à la séquence suivante. Les résultats ont montré que 2 à 3 visionnages de la séquence étaient en moyenne nécessaires au sujet pour arriver à une balance satisfaisante. Chaque séquence était initialement présentée avec une valeur aléatoire pour n, afin d'éviter une éventuelle influence de la balance initiale fixée par

les expérimentateurs. La diffusion des stimuli et la récupération des données de l'interface MIDI étaient assurées par un logiciel programmé sous Max/MSP.

Sujets

11 sujets (3 femmes et 8 hommes, âgés de 20 à 25 ans) ont participé à ce test. Il s'agissait d'étudiants en formation aux métiers du son et de l'image (Master Image & Son de l'Université de Brest). Bien que n'ayant pas d'expérience des tests perceptifs, ces étudiants sont formés au mixage de contenus audiovisuels.

4.2.2 Résultats

Une analyse statistique a été effectuée sur les données, en prenant ΔG comme variable, pour déterminer s'il y avait des différences de balances significatives entre mixages 3D-s (c'est-à-dire mixages avec projection des images en 3D-s) et mixages 2D (c'est-à-dire mixages avec projection des images en 2D) et une éventuelle interaction avec les séquences.

On utilise souvent pour ce genre de tests une analyse de variance (ANOVA). Plusieurs hypothèses doivent cependant être vérifiées pour que son utilisation soit légitime (Howell, 2009). L'hypothèse de normalité, par exemple, exige que les observations de chaque cellule (une cellule est un groupe d'observations appartenant à la même combinaison de variables indépendantes) suivent une distribution normale. Un test de Kolmogorov-Smirnov (niveau de significativité à 5%) a rejeté l'hypothèse d'une distribution normale pour 19 cellules sur 44. La plupart des études statistiques affirment que l'ANOVA peut être robuste à ce genre de violations si la taille d'échantillon est suffisamment large (supérieure à 15 observations par cellule selon Green et Salkind (2013)). Avec 11 observations par cellule, il est donc préférable d'utiliser un test non-paramétrique adapté aux mesures répétées : le test de Wilcoxon (Hollander et Wolfe, 1999).

Les résultats, comparant les mixages 3D-s et 2D pour chaque séquence, sont présentés dans le Tableau 4.1. Comme les tests non-paramétriques présentent l'avantage d'être robustes à la présence d'outliers, nous avons décidé dans un premier temps de garder le sujet 6 dans nos analyses.

Durant la première session du test, seule la séquence 10 a donné lieu à des mixages différents en 2D et en 3D-s ($p = 0.033$). Pour cette séquence, les mixages étaient plus frontaux lorsque l'image était projetée en 3D-s que lorsqu'elle était projetée en 2D.

Une fois la première session terminée, le sujet prenait une pause de 15 minutes puis passait à nouveau le test. L'analyse de la seconde session montre des différences significatives pour 3 séquences : la séquence 2 ($p = 0.013$), la séquence 4 ($p = 0.047$) et la séquence 9 ($p = 0.021$). Pour ces trois séquences, les mixages 2D étaient cette fois-ci plus frontaux que les mixages 3D-s. Les autres séquences n'ont pas donné lieu à des balances significativement différentes,

Tableau 4.1 – Résultats du test de Wilcoxon pour l'expérience I, comparant les mixages 2D et 3D-s pour chaque session et chaque séquence.

Session	Séq.	Z	Sig. P	Rang
1	1	-0.267	0.79	Diff. non significative
1	2	-0.089	0.929	Diff. non significative
1	3	-1.6	0.11	Diff. non significative
1	4	-0.178	0.859	Diff. non significative
1	5	-0.089	0.929	Diff. non significative
1	6	-1.479	0.139	Diff. non significative
1	7	-0.711	0.477	Diff. non significative
1	8	-1.067	0.286	Diff. non significative
1	9	-0.089	0.929	Diff. non significative
1	10	-2.134	0.033	**mixage 3D-s plus frontal**
1	11	-0.561	0.575	Diff. non significative
2	1	-1.067	0.286	Diff. non significative
2	2	-2.497	0.013	**mixage 2D plus frontal**
2	3	-1.423	0.155	Diff. non significative
2	4	-1.988	0.047	**mixage 2D plus frontal**
2	5	-0.8	0.424	Diff. non significative
2	6	-1.334	0.182	Diff. non significative
2	7	-0.153	0.878	Diff. non significative
2	8	-0.533	0.594	Diff. non significative
2	9	-2.312	0.021	**mixage 2D plus frontal**
2	10	-1.58	0.114	Diff. non significative
2	11	-0.089	0.929	Diff. non significative

séquence 10 incluse (p = 0.114).

La session semble donc avoir eu une influence significative sur les résultats. Les Fig. 4.2, 4.3, 4.4 et 4.5 montrent pour chaque sujet les résultats obtenus dans les sessions 1 et 2 pour les séquences 2, 4, 9 et 10 (c'est-à-dire les séquences qui ont donné lieu, soit dans la première session, soit dans la deuxième session, à des différences de mixages significatives entre 2D et 3D-s).

Séquence 2

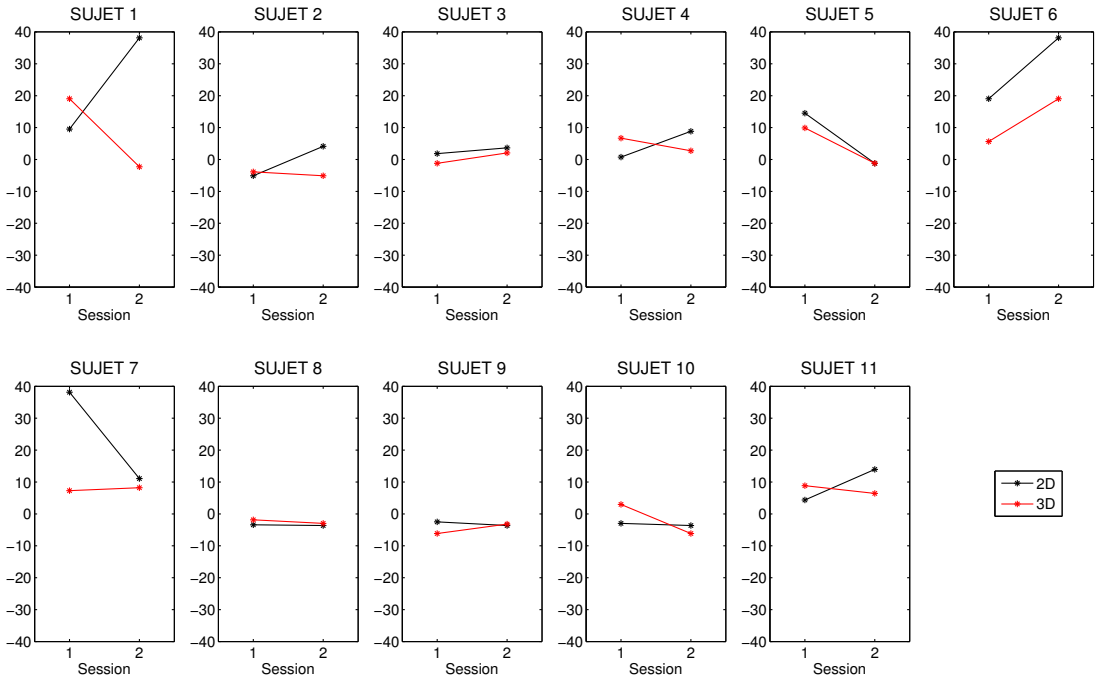


FIGURE 4.2 – Comparaison des balances obtenues dans les sessions 1 et 2, pour chaque sujet, dans la séquence 2

Pour la séquence 2, il n'y a pas eu de différence significative dans la session 1. Par contre, dans la session 2, le test de Wilcoxon indique que les mixages avec images 3D ont été plus « surround » que les mixages avec images 2D, en accord avec notre sous-hypothèse 1. La Fig. 4.2 montre que cet effet est dû à huit sujets. Les trois sujets restants (5, 8 et 9) ont produit des balances quasi-identiques en 2D et en 3D-s.

En regardant sujet par sujet les résultats obtenus pour la séquence 2, nous pouvons identifier plusieurs types d'évolution d'une session à l'autre :

— certains sujets présentent un « croisement » d'une session à l'autre. Par exemple, parmi les huit sujets qui ont produit des mixages 2D plus frontaux que les mixages 3D-s dans la session 2, cinq sujets avaient mixé « dans l'autre sens » lors de la première session (sujets 1, 2, 4, 10 et 11), c'est-à-dire avec des mixages 3D-s plus frontaux que les mixages 2D ;

— certains sujets sont restés « cohérents » d'une session à l'autre. En effet, les sujets 3,

6 et et 7 ont toujours produit un mixage plus frontal pour la version 2D que pour la version 3D-s de la séquence 2, aussi bien dans la première que dans la seconde session. Cette cohérence est cependant relative : par exemple, le sujet 6 a globalement produit des mixages beaucoup plus frontaux lors de la seconde session, aussi bien en 3D-s qu'en 2D. Quant au sujet 7, le mixage de la version 2D a été beaucoup moins frontal lors de la seconde session.

.

Séquence 4

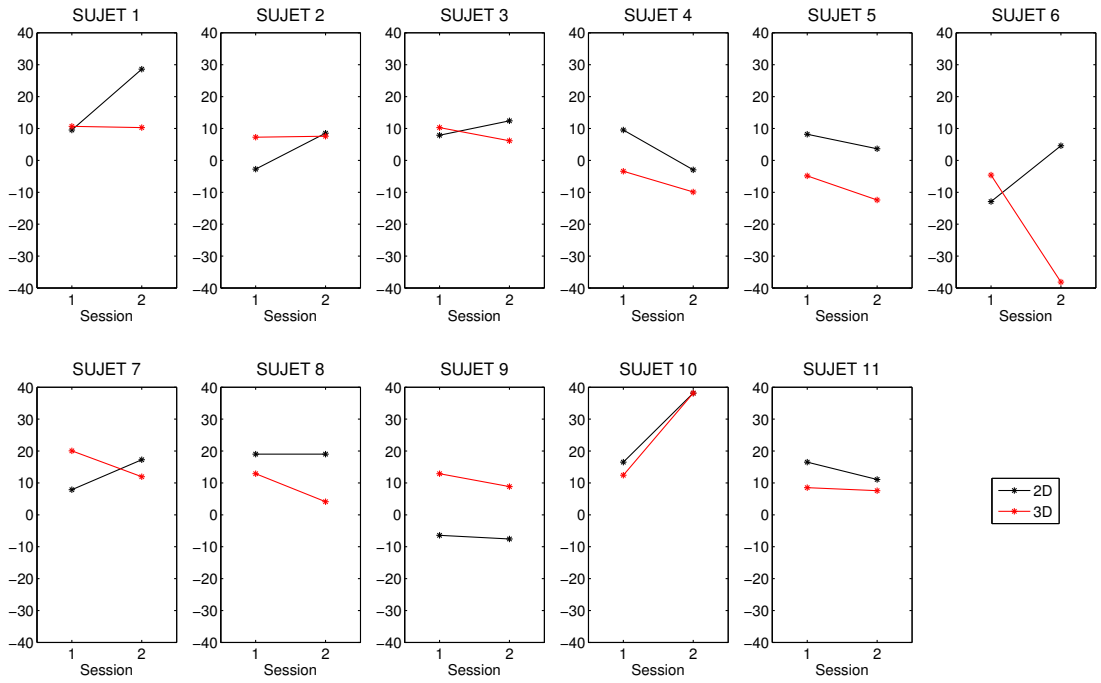


FIGURE 4.3 – Comparaison des balances obtenues dans les sessions 1 et 2, pour chaque sujet, dans la séquence 4

Pour la séquence 4, il n'y a pas eu de différence significative dans la session 1. Par contre, dans la session 2, le test de Wilcoxon indique que les mixages avec images 3D ont été globalement plus « surround » que les mixages avec images 2D, en accord avec notre sous-hypothèse 1. La Fig. 4.3 montre que cet effet est dû à neuf sujets (dont le sujet 2 pour lequel la différence de balance est cependant très faible). Parmi les deux sujets restants, le sujet 9 présente une tendance inverse (mixage avec images 2D plus « surround » que le mixage avec images 3D) et le sujet 10 a fixé la balance de la même manière en 2D et en 3D-s.

En regardant sujet par sujet les résultats obtenus pour la séquence 4, nous remarquons à nouveau un nombre important de « croisements » d'une session à l'autre (chez les sujets 1, 2, 3, 6, et 7). Même pour les sujets qui n'ont pas « croisé » (sujets 4, 5, 8, 9 et 11), les mixages ont été globalement plus « surround » dans la seconde session que dans la première session.

Séquence 9

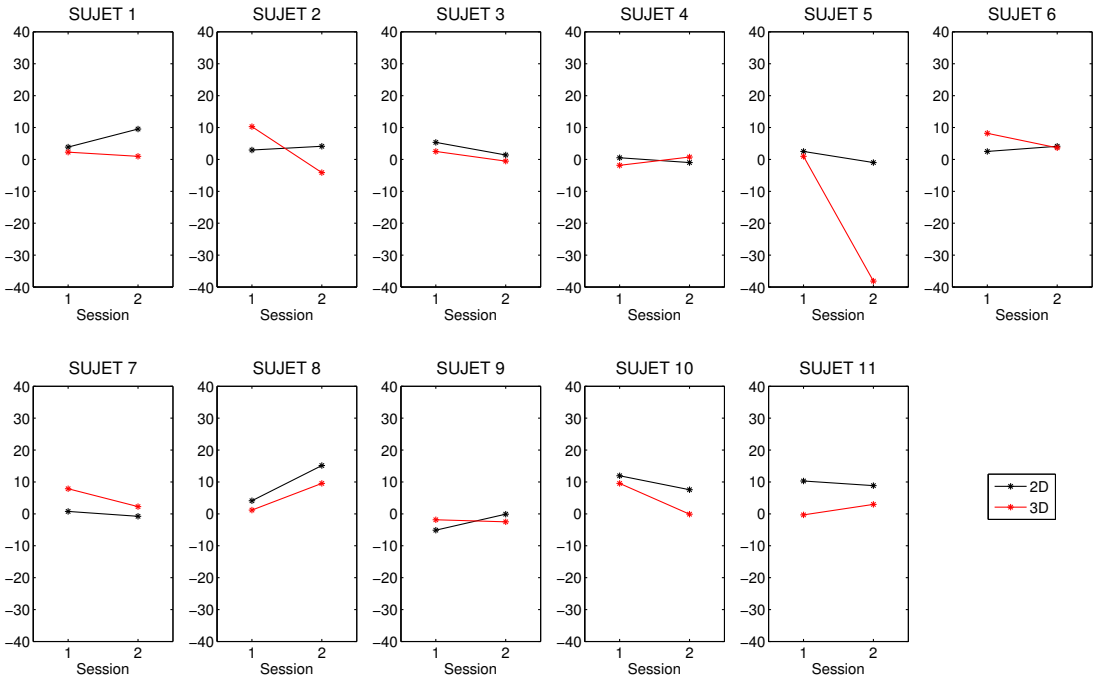


FIGURE 4.4 – Comparaison des balances obtenues dans les sessions 1 et 2, pour chaque sujet, dans la séquence 9

Pour la séquence 9, il n'y a pas eu de différence significative dans la session 1. Par contre, dans la session 2, le test de Wilcoxon indique que les mixages avec images 3D ont été globalement plus « surround » que les mixages avec images 2D, en accord avec notre sous-hypothèse 1. La Fig. 4.4 montre que cet effet est dû à neuf sujets (dont le sujet 6 pour lequel la différence de balance est cependant très faible). Les deux sujets restants (sujets 4 et 7) présentent une tendance inverse (mixage avec images 2D plus « surround » que le mixage avec images 3D).

Nous constatons à nouveau que parmi les neuf sujets ayant produit un mixage plus surround avec images 3D-s dans le seconde session, quatre sujets présentaient une tendance inverse dans la première session. Parmi les sujets qui ne « croisent » pas, les mixages peuvent quand même être différents d'une session à l'autre. Par exemple, les mixages du sujet 8 sont globalement plus frontaux dans la seconde session, que l'image soit en 2D ou en 3D-s.

Séquence 10

Pour la séquence 10, il n'y a eu de différence significative que pour la session 1, dans laquelle le test de Wilcoxon indique que les mixages avec images 2D ont été globalement plus « surround » que les mixages avec images 3D, contrairement à notre sous-hypothèse 1. La Fig. 4.5 montre que cet effet est dû à huit sujets. Les trois sujets restants (sujets 7, 9 et 10) présentent une tendance inverse (mixage avec images 3D plus « surround » que le mixage avec images 2D).

Parmi les huit sujets ayant produit un mixage plus frontal avec images 3D-s dans le

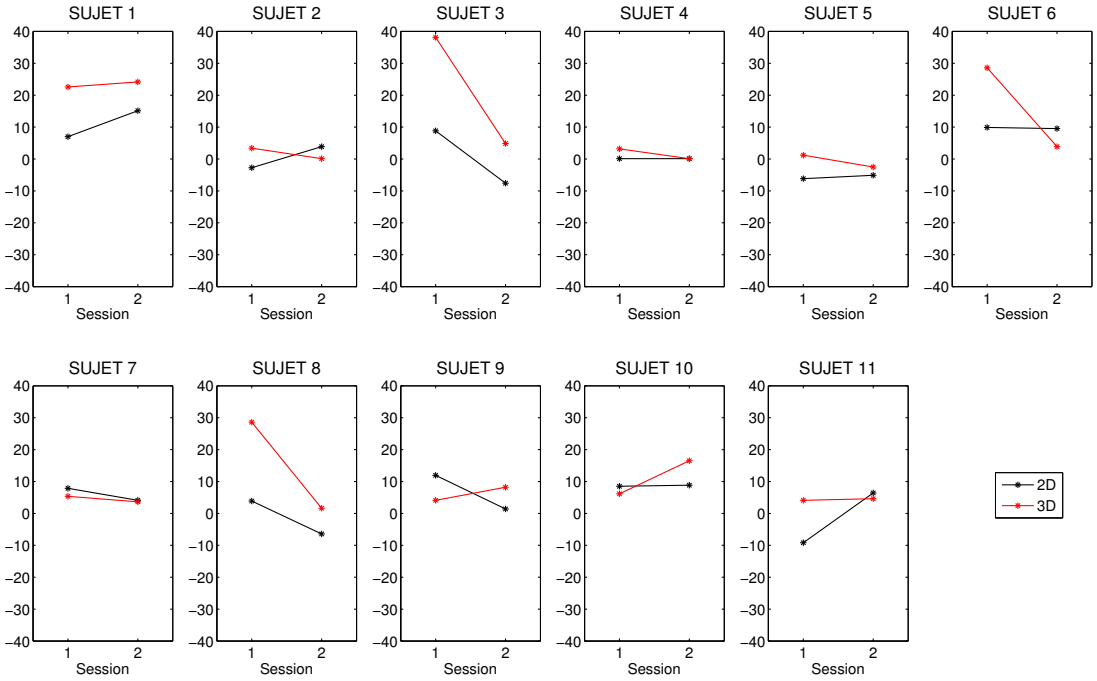


FIGURE 4.5 – Comparaison des balances obtenues dans les sessions 1 et 2, pour chaque sujet, dans la séquence 10

première session, trois sujets ont finalement « changé d'avis » dans la seconde session. Parmi les sujets qui ne « croisent » pas, les mixages peuvent quand même être différents d'une session à l'autre. Par exemple, les mixages du sujet 8 sont cette fois-ci globalement plus « surround » dans la seconde session, que l'image soit en 2D ou en 3D-s.

Nous constatons donc pour les séquences 2, 4, 9 et 10 une certaine instabilité d'une session à l'autre pour une grande partie des sujets.

Recherche de corrélations entre les tailles des boîtes scéniques et les différences de mixages

Le tableau 4.2 montre les tailles des boîtes scéniques mesurées pour les versions 3D-s de chaque séquence. Ces tailles sont déterminées en mesurant en pixels la différence entre la disparité de l'objet le plus loin et la disparité de l'objet le plus proche. Ces mesures permettent de décrire objectivement la « quantité » de stéréoscopie dans une séquence, mais elles ne permettent pas toujours de prédire correctement le degré de sensation de profondeur visuelle perçue par les sujets (Moulin, 2015).

Notre hypothèse est que plus la boîte scénique d'une séquence est grande, plus la différence entre mixages 2D et 3D-s est marquée. Une lecture rapide du tableau montre que, certes, la séquence 2 a la boîte scénique la plus grande, cependant les séquences 4, 9 et 10 n'émergent pas spécialement au-dessus des autres séquences. Une analyse montre même que les tailles de boîtes scéniques ne sont corrélées avec les différences entre mixages 2D et 3D-s pour aucun sujet dans aucune session.

Séquence	boîte scénique (px)
1	de 8 à 63
2	47
3	27
4	31
5	32
6	31
7	26
8	16
9	27
10	24
11	31

Tableau 4.2 – Boîtes scéniques pour chaque séquence de l'expérience I.

Bilan

Nous constatons donc, pour les quatre séquences ayant donné lieu à des mixages significativement différents en 2D et en 3D-s, une grande instabilité d'une session à l'autre. Les différences significatives détectées par le test de Wilcoxon pour certaines séquences dans certaines sessions sont en partie dues à des sujets qui ont présenté des tendances totalement inverses d'une session à l'autre. Plusieurs facteurs peuvent expliquer cette instabilité :

— les importantes différences constatées d'une session à l'autre pour chaque sujet sont peut-être dues à l'effet d'ordre des stimuli, que nous avons ici tenté de minimiser en rendant l'ordre aléatoire et différent pour chaque sujet ;
— les différences de résultats entre les deux sessions traduisent peut-être un effet d'apprentissage de la tâche. Cette hypothèse permettrait d'expliquer pourquoi plus de séquences ont été significatives dans la seconde session que dans la première session ;
— la tâche est peut-être trop compliquée ou le niveau d'expertise des sujets trop faible. Dans ce cas, il faut prévoir de reconduire l'expérience avec des sujets plus expérimentés (mixeurs professionnels avec plusieurs années d'expérience) ou alors organiser une nouvelle expérience avec une tâche plus simple.

Nous avions remarqué que le sujet 6 avait été un « outlier » pour 8 conditions du test, dont les séquences 2 et 4 lors de la seconde session. En retirant ce sujet des données, nous observons une augmentation des niveaux de significativité p associés à chaque comparaison de mixages 3D-s et 2D, ce qui entraine un léger dépassement du seuil à 0.05 pour la séquence 4 session 2 et la séquence 10 session 1, dont les valeurs de p étaient déjà importantes à la base ($p = 0.047$ et $p = 0.033$ respectivement). Les mixages 3D-s et 2D des séquences 4 et 10 ne peuvent donc plus, en toute rigueur, être considérés comme significativement différents dès lors que les résultats du sujet 6 sont retirés des données (une décision qui est cependant discutable). Dans tous les cas, le nombre de différences significatives observées est faible.

4.3 Expérience II : Influence de la stéréoscopie sur la perception de la balance frontal/surround de sons d'ambiance

Dans l'expérience I, plus de différences significatives ont été obtenues dans la seconde session que dans la première session. Ce constat suggère que les sujets ont peut-être dû passer par une phase d'apprentissage avant de pouvoir donner des résultats pertinents, ou alors que la tâche était trop compliquée pour que les sujets parviennent à rester cohérents dans leurs mixages. De plus, le nombre de sujets était faible.

Une seconde expérience a donc été organisée, avec plus de sujets, une session supplémentaire (donc 3 sessions en tout) et une tâche simplifiée. Le test a cette fois-ci eu lieu dans un cinéma, avec des contenus 3D-s issus de productions professionnelles. Les sujets ne pouvaient pas fixer eux-même leur propre balance frontal/surround et devaient plutôt évaluer des balances fixées au préalable par les expérimentateurs.

4.3.1 Matériel et méthode

Lieu de l'expérience

Le test s'est déroulé dans un cinéma 5.1 de grande taille (une photo de la salle est donnée en Fig. 4.6, et un plan détaillé en Fig. 4.7), avec :

- 3 enceintes à 2 voies passives KCS S-2500 pour les canaux frontaux : Gauche, Centre et Droite ;
- 12 enceintes à 2 voies passives KCS SR-15 pour les enceintes surround, divisées en deux canaux gauche et droit ;
- 1 caisson de basse KCS C-118-A.

Cette configuration présente une différence majeure par rapport à celle de l'expérience I. En effet, il n'y avait qu'une enceinte par canal surround dans l'expérience I alors qu'il y a ici 6 enceintes pour le canal surround gauche et 6 enceintes pour le canal surround droit.

FIGURE 4.6 – Photo du cinéma utilisé dans le cadre de l'expérience II.

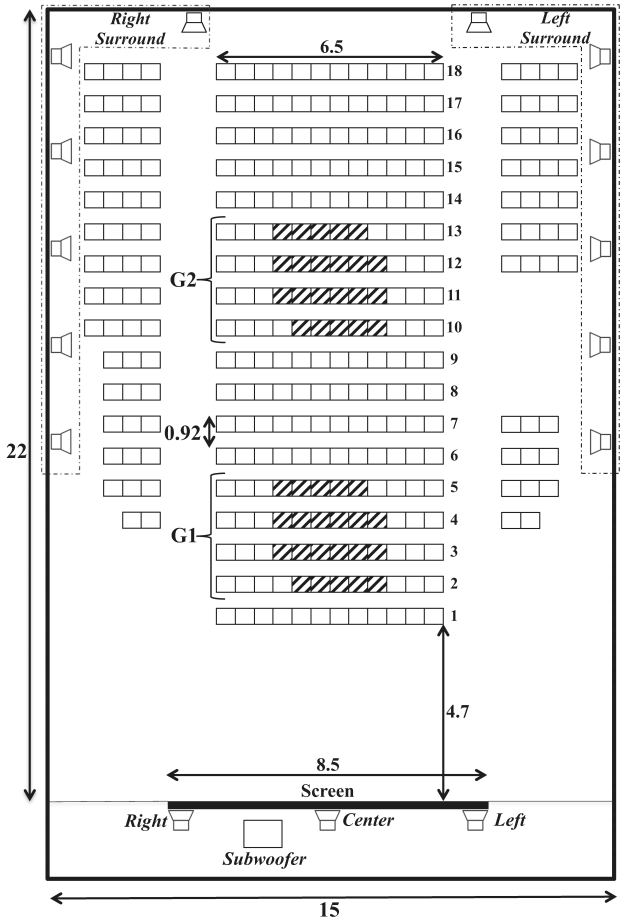


FIGURE 4.7 – Plan du cinéma pour l'expérience II. Les sièges occupés par les sujets sont grisés. Dimensions en mètre.

Les enceintes étaient alimentées par un processeur numérique Dolby CP750 et des amplificateurs QSC USA900. Les gains avaient été ajustés selon les recommandations Dolby, de telle sorte que :

— chaque enceinte frontale produise individuellement au point d'écoute de référence (c'est-à-dire au deux-tiers de la salle en partant de l'écran, aligné sur le centre de l'écran) un niveau de pression sonore égal à 85 dBC, lorsqu'est diffusé un bruit rose à - 20 dBFS RMS ;
— chaque canal surround (enceintes surround gauches d'une part, enceintes surround droites d'autre part) produise individuellement un niveau de pression sonore égal à 82 dBC ;
— le canal LFE produise un niveau de pression sonore égal à 91 dBC.

La fréquence d'échantillonnage des signaux audio était de 48 kHz et la quantification était effectuée sur 24 bits.

L'image (24 i/s) était projetée par un projecteur numérique Christie CP2000-ZX, synchronisé avec des lunettes 3D-s actives Eyes3Shut Purple Two.

Séquences

La première expérience a montré que l'effet de la stéréoscopie dépendait du contenu des séquences. Il a donc été décidé pour ce deuxième test de garder le plus de séquences possible. Cependant, pour que la durée du test reste raisonnable, le nombre de séquences a dû être réduit à 8.

5 séquences extraites de productions professionnelles ont été utilisées, ainsi que trois séquences issues de l'expérience préliminaire :

— le dialogue dans un café ;

— deux marins effectuant une manœuvre à la proue d'un bateau ;

— la foule dans une rue.

Il est à noter que les expérimentateurs avaient assuré la prise de son, le montage et le mixage de la bande-son originale de 3 des 5 séquences extraites de productions professionnelles (voir Fig. 4.8). Les séquences sont plus largement détaillées dans l'annexe E.

FIGURE 4.8 – Photo du tournage de la séquence 8 de l'expérience II.

Echelle de jugement et protocole

Des questionnaires papier ont été distribués aux sujets. Pour chaque stimulus leur était proposée une échelle graduée, dont les extrémités étaient labellisées « beaucoup trop frontal » et « beaucoup trop « surround » » (voir Fig. 4.9).

Le sujet devait choisir une des gradations parmi les 11 proposées en l'entourant. Il n'y avait aucun label intermédiaire prédéfini, pour éviter l'introduction de biais indésirable (Poulton, 1992; Zielinski *et al.*, 2008). Les 11 gradations ont été par la suite converties en valeurs numériques (de 0 à 10) pour l'analyse statistique (avec les extrémités « beaucoup trop « surround » » et « beaucoup trop frontal » respectivement égales à 0 et 10).

L'échelle était divisée en 11 gradations séparées par des intervalles égaux, ce qui signifie que l'échelle peut être considérée comme continue (Nunnally et Bernstein, 1994) et les données peuvent donc être analysées à l'aide de procédures paramétriques, tant que certaines hypothèses ne sont pas violées trop sévèrement (Bech et Zacharov, 2006).

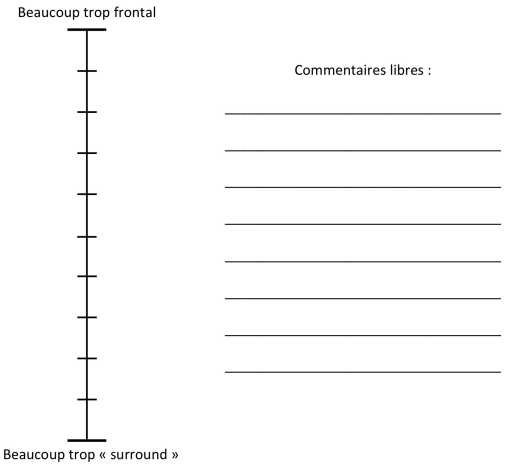


FIGURE 4.9 – Échelle de jugement pour chaque stimulus lors de l'expérience II

Les sujets étaient priés de garder leurs lunettes 3D-s pendant la totalité de l'expérience, même pour les séquences 2D (durant lesquelles la même image était envoyée à l'œil gauche et à l'œil droit), afin d'éviter une éventuelle influence de la perte de luminosité.

Les stimuli duraient environ 20 secondes chacun. Après la présentation d'un stimulus, des lumières tamisées s'allumaient pour permettre aux sujets de reporter leur réponse sur la feuille. Au bout de 20 secondes, les lumières s'éteignaient et le stimulus suivant était présenté. Tous les sujets ont passé le test en même temps, sans pause entre les sessions pour éviter que les sujets ne se parlent entre eux et ne s'influencent mutuellement.

Balances sonores

Pour chaque séquence, deux mixages ont été produits par les expérimentateurs, de telle sorte qu'un des mixages (mixage A) « sonne » substantiellement plus frontal que l'autre (mixage B). Une analyse a montré que :

- les enceintes frontales étaient en moyenne 2 dB plus fortes dans le mixage A que dans le mixage B ;
- les enceintes surround étaient en moyenne 5.3 dB moins fortes dans le mixage A que dans le mixage B.

Si la décision de diffuser chaque séquence avec deux mixages différents présente le désavantage de multiplier par 2 la durée du test, elle permet néanmoins de :

- réduire l'influence des expérimentateurs, qui fixent les balances frontal/surround subjectivement ;
- éviter un effet de compression des données : si un seul mixage avait été proposé aux sujets et que ce dernier avait été dans l'absolu trop frontal (ou trop « surround »), les sujets n'auraient probablement exploité que le haut (ou le bas) de l'échelle de jugement, ce qui aurait pu minimiser voire masquer des différences significatives entre versions 2D et 3D-s ;

— étudier d'éventuelles interactions entre le Mode Visuel et la Balance frontal/surround proposée.

Pour chaque séquence, les niveaux globaux d'écoute ont été égalisés subjectivement par les expérimentateurs (une pratique courante dans les tests subjectifs (Standard AES20-1996 (r2007), 2008)), de manière à ce que les mixages A soient perçus au même niveau que les mixages B. Le niveau global d'écoute de chaque séquence a également été fixé subjectivement, comme c'est souvent le cas dans les salles de cinéma (Recommandation RT 013, 2006).

Chaque séquence a donc été présentée sous 4 modalités différentes aux sujets :

— avec l'image 3D-s et le mixage A ;
— avec l'image 2D et le mixage A ;
— avec l'image 3D-s et le mixage B ;
— avec l'image 2D et le mixage B.

Sujets

44 personnes (10 femmes et 34 hommes, âgés de 20 à 25 ans) ont pris part à l'expérience. Il s'agissait d'étudiants en formation aux métiers du son et de l'image (Master Image & Son de l'Université de Brest). Bien que n'ayant pas d'expérience des tests perceptifs, ces étudiants sont formés à l'écoute analytique et au mixage de contenus audiovisuels.

Groupes

La perception de la balance entre enceintes frontales et surround dépend directement du placement du sujet dans la longueur de la salle (Toole, 2008). Il a donc été décidé de séparer les sujets en 2 groupes :

— un premier groupe en proximité de l'écran (rangées n°2, 3, 4 et 5) ;
— un second groupe au « Sweet Spot » (rangées n°10, 11, 12 et 13),

l'effet de la rangée étant supposé négligeable à l'intérieur d'un groupe. Avec 22 sujets par groupe, répartis sur 4 rangées, les sujets étaient assis à une distance maximale de 1,35 mètre du centre de la rangée. Cette distance étant faible par rapport à la taille de la salle (22 mètres de long pour 15 mètres de large), son impact sur la perception des niveaux et des différences interaurales gauche/droite devrait être négligeable (Toole, 2008).

La perception spatiale a certainement été très différente d'un groupe à l'autre : comme le montre la Fig. 4.7, toutes les enceintes surround se trouvaient derrière les sujets pour le premier groupe en proximité de l'écran. Par contre, pour le second groupe au « Sweet Spot », les enceintes surround se répartissaient devant, sur les côtés et derrière les sujets. Les balances étaient également plus frontales pour les sujets du groupe 1, puisqu'ils étaient plus proches des enceintes frontales et plus éloignés des enceintes surround que les sujets du groupe 2.

La perception visuelle a également dû être très différente : pour le premier groupe en proximité de l'écran, le champ visuel de l'écran correspondait à un angle d'environ 62.5° alors que celui du second groupe au « Sweet Spot » correspondait à un angle de 33°.

Répétition

La première expérience avait montré une influence de la session. Il a donc été décidé de répéter le test trois fois de suite. Les sujets avaient donc 8 séquences × 2 modes visuels (2D et 3D-s) × 2 balances (A et B) × 3 sessions = 96 présentations à évaluer, soit une heure de test au total.

4.3.2 Résultats

Pour minimiser les différences inter-sujets d'utilisation de l'échelle, les données brutes ont subi une transformée en z (ITU-R BS.1116-1, 1994; ITU-R BS.1286, 1997), de telle sorte que :

$$Z_i = \frac{(x_i - x_{si})}{s_{si}} \cdot s_s + x_s$$

avec :

— Z_i : résultat normalisé ;

— x_i : note donnée par le résultat du participant i ;

— x_{si} : note moyenne du participant i pendant la séance s ;

— x_s : note moyenne de tous les participants pendant la séance s ;

— s_s : écart type pour tous les participants pendant la séance s ;

— s_{si} : écart type pour le participant i pendant la séance s.

Vérification des hypothèses pour une ANOVA

Plusieurs hypothèses doivent être validées pour pouvoir utiliser une ANOVA mixte légitimement (Howell, 2009) :

— l'hypothèse de normalité ;

— l'hypothèse d'homoscédasticité ;

— l'hypothèse de sphéricité.

Comme dans la première expérience, un test de Kolmogorov-Smirnov a été effectué sur chacune des cellules (niveau de significativité à 5%) pour vérifier l'hypothèse de normalité. Le test a rejeté l'hypothèse d'une distribution normale pour 12 cellules sur 192. Cependant, la taille d'échantillon est supérieure à 15 observations par cellule (Green et Salkind, 2013), et de nombreuses études ont recours à l'ANOVA même lorsque la plupart des cellules dévient significativement de la normalité (Zielinski *et al.*, 2003). Avec 22 observations par cellule, et seulement 12 cellules sur 192 violant l'hypothèse de normalité, l'utilisation d'une ANOVA reste envisageable.

L'hypothèse d'homoscédasticité signifie que les variances obtenues pour chaque combinaison de variables indépendantes intra-sujets doivent être les mêmes d'un groupe à l'autre. Un test de Levene a montré que cette hypothèse était violée pour 10 combinaisons sur 96. Cependant, l'ANOVA peut encore donner des résultas pertinents si le nombre d'observations dans chaque cellule est le même (Zielinski *et al.*, 2003), ce qui est le cas dans notre étude.

Comme il s'agit d'une ANOVA mixte, l'analyse intègre des mesures répétées et une hypothèse supplémentaire doit donc être vérifiée : l'hypothèse de sphéricité, à savoir que les variances des différences entre toutes les paires possibles de combinaisons de variables indépendantes intra-sujets doivent être égales. Un test de Mauchly a montré que cette hypothèse était violée pour les facteurs « Séquence », « Séquence * Répétition » et « Séquence * Répétition * Mode Visuel ». La valeur F de ces facteurs a donc dû être corrigée, en utilisant soit la correction de Huynh-Feldt, soit celle de Greenhouse-Geisser (Girden, 1991). Une ANOVA mixte peut donc être légitimement utilisée pour l'analyse des données.

ANOVA

Les résultats de l'analyse de variance sont présentés dans le Tableau 4.3, avec les effets des 5 facteurs suivants :

Intra-sujets

- S : Séquence (8 niveaux) ;
- V : Mode Visuel (2 niveaux : 2D vs. 3D-s) ;
- B : Balance Sonore (2 niveaux : mixage A vs. mixage B) ;
- R : Répétition (3 niveaux).

Inter-sujets

- G : Groupe (2 niveaux : groupe 1 « Proche de l'écran » vs. groupe 2 « au Sweet Spot »).

Influence du Mode Visuel : 3D-s vs. 2D

Les résultats montrent une influence significative du Mode Visuel sur les notes des sujets ($V : F(1, 42) = 9.566$, $p = 0.004 < 0.01$) avec une valeur moyenne de 5.54 pour les balances avec images 3D-s et de 5.36 pour les balances avec images 2D, ce qui veut dire que les sujets ont globalement perçu les mixages avec images 3D-s comme étant plus frontaux que ceux avec images 2D.

L'interaction Séquence - Mode Visuel ($S*V : F(7, 294) = 2.38$, $p = 0.022 < 0.05$) est également significative, ce qui veut dire que l'effet du Mode Visuel n'a pas été le même selon la séquence. Un *post-hoc* LSD de Tukey, comparant l'effet du Mode Visuel séquence par séquence (Fig. 4.10), montre que le Mode Visuel n'a eu finalement d'impact que pour les séquences 1 (dialogue dans un café, $p = 0.001$) et 2 (marins effectuant des manœuvres sur un

Tableau 4.3 – Résultats de l'ANOVA pour l'expérience II.

Source	SS	DF	MS	F	Sig. p
S	743.993	6.698	111.081	21.356	< 0.001
V	33.375	1	33.375	9.566	0.004
B	691.159	1	691.159	156.687	< 0.001
R	22.596	2	11.298	3.145	0.048
G	30.004	1	30.004	4.40E+16	< 0.001
S*V	21.325	7	3.046	2.38	0.022
S*B	211.016	7	30.145	16.475	< 0.001
S*R	64.012	9.335	6.857	3.25	0.001
S*G	184.527	7	26.361	5.297	< 0.001
V*B	0.487	1	0.487	0.445	0.509
V*R	8.417	2	4.208	4.08	0.02
V*G	22.571	1	22.571	6.469	0.015
B*R	27.362	2	13.681	14.604	< 0.001
B*G	57.836	1	57.836	13.111	0.001
R*G	0.445	2	0.222	0.062	0.94
S*V*B	12.912	7	1.845	1.627	0.127
S*V*R	37.491	9.728	3.854	1.925	0.042
S*V*G	13.778	7	1.968	1.538	0.154
S*B*R	35.855	14	2.561	1.932	0.021
S*B*G	28.922	7	4.132	2.258	0.03
S*R*G	11.436	14	0.817	0.581	0.881
V*B*R	0.021	2	0.011	0.011	0.989
V*B*G	2.917	1	2.917	2.663	0.11
V*R*G	12.366	2	6.183	5.994	0.004
B*R*G	2.555	2	1.277	1.363	0.261
S*V*B*G	24.181	7	3.454	3.047	0.004
S*V*R*G	30.037	14	2.146	1.542	0.091
S*V*B*R	45.138	14	3.224	2.532	0.002
S*B*R*G	14.718	14	1.051	0.793	0.677
V*B*R*G	0.836	2	0.418	0.447	0.641
S*V*B*R*G	19.249	14	1.375	1.08	0.373

bateau, $p = 0.003$).

Les notes moyennes avec leur intervalle de confiance à 95%, obtenues pour le Mode Visuel séquence par séquence, sont présentées dans la Fig. 4.10.

Influence de la Répétition

Les résultats indiquent une influence de la répétition (R : $F(2, 84) = 3.145$, $p = 0.048 < 0.05$), avec pour les 3 sessions des moyennes successives de 5,517, 5,486, et 5,348.

L'interaction Mode Visuel - Répétition est également significative (V*R : $F(2, 84) = 4.08$, $p = 0.02 < 0.05$). Un *post-hoc* LSD de Tukey montre que l'effet du Mode Visuel a été significatif pour les sessions 1 ($p < 0.001$) et 2 ($p = 0.024$) mais ne l'était plus lors de la troisième et dernière session ($p = 0.276$, Fig. 4.11).

Les mixages avec images 2D n'ont pas été significativement différents entre eux d'une

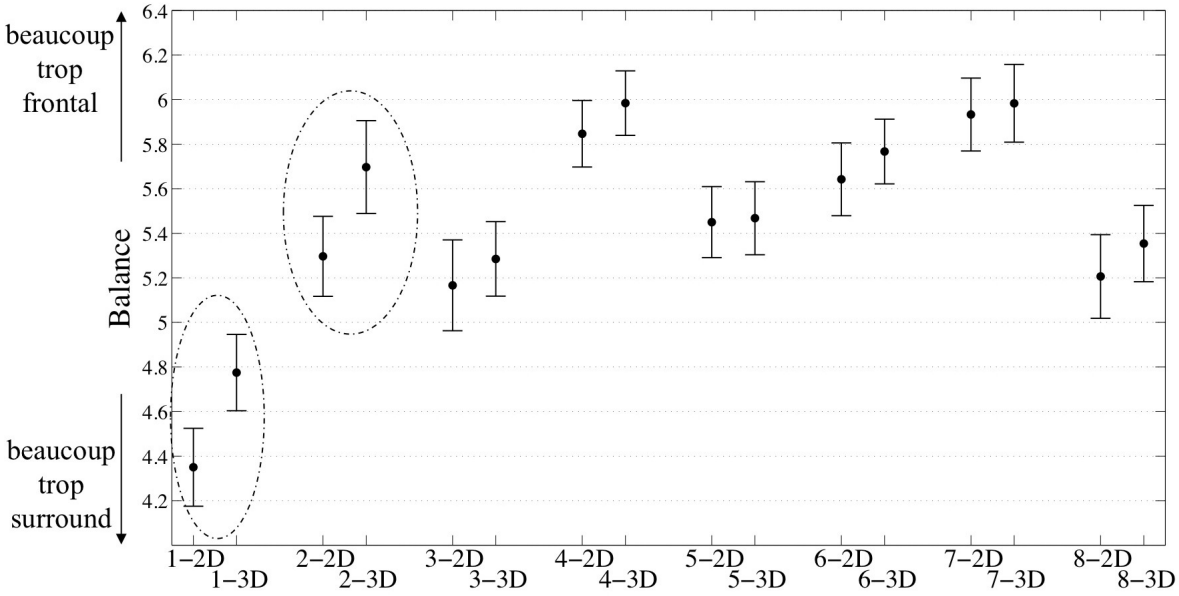


FIGURE 4.10 – Effet du Mode Visuel (2D vs. 3D-s) sur la perception de la balance frontal/surround en fonction de la Séquence lors de l'expérience II. Moyennes et intervalles de confiance à 95%.

session à l'autre (p = 1.000 entre les sessions 1 et 2, p = 0.691 entre les sessions 2 et 3). Les mixages avec images 3D n'ont pas été significativement différents entre eux entre les sessions 1 et 2 (p = 0.214) mais l'ont été entre les sessions 2 et 3 (p = 0.016) . Le fait qu'il n'y ait plus eu de différence significative entre mixages 2D et mixages 3D-s lors de la session 3 est donc dû à une décroissance significative des notes attribuées aux mixages 3D-s (p = 0.049) entre les sessions 2 et 3.

Les notes moyennes avec leur intervalle de confiance à 95%, obtenues pour le Mode Visuel session par session, sont présentées dans la Fig. 4.11.

Influence du Groupe : Proximité de l'écran vs. « Sweet Spot »

Les résultats indiquent une influence significative du groupe (G : F(1, 42) = 4.4E+16, p < 0.001) avec un valeur moyenne de 5.535 pour le groupe 1 et une valeur moyenne de 5.366 pour le groupe 2.

Les sujets assis en proximité de l'écran ont donc globalement perçu les mixages comme étant plus frontaux que les sujets assis au « Sweet Spot », ce qui n'est pas surprenant puisque le groupe 1 était plus proche des enceintes frontales que le groupe 2.

L'interaction entre le Mode Visuel et le Groupe (V*G : F(1, 42) = 6.469, p = 0.015 < 0.05) est également significative. Un *post-hoc* LSD de Tukey montre qu'il n'y a finalement eu d'effet significatif du Mode Visuel que pour le groupe 2 (Fig. 4.12), soit le groupe assis au « Sweet Spot » (p < 0.001), tandis qu'il n'y en a pas eu pour le groupe 1, assis en proximité de l'écran (p = 0.700).

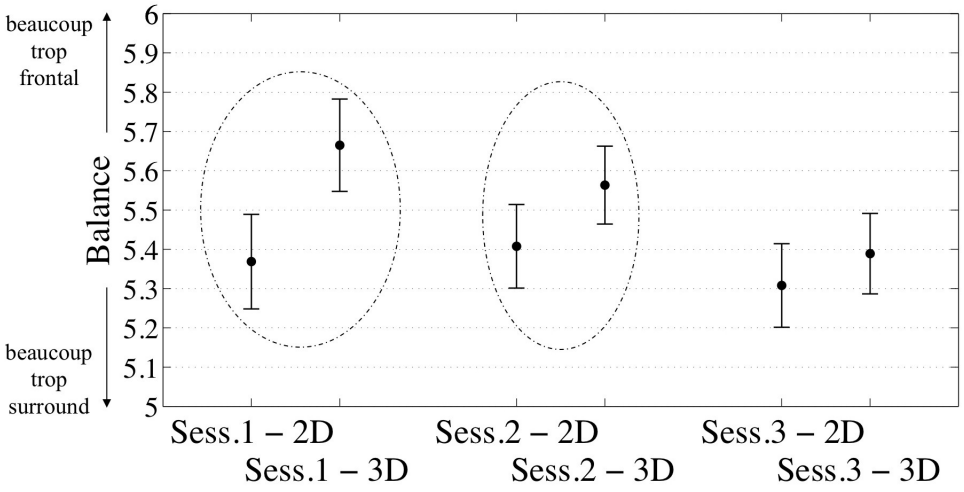


FIGURE 4.11 – Effet du Mode Visuel (2D vs. 3D-s) sur la perception de la balance frontal/surround en fonction de la Session lors de l'expérience II. Moyennes et intervalles de confiance à 95%.

Les notes moyennes avec leur intervalle de confiance à 95%, obtenues pour le Mode Visuel pour chacun des 2 groupes, sont présentées dans la Fig. 4.12.

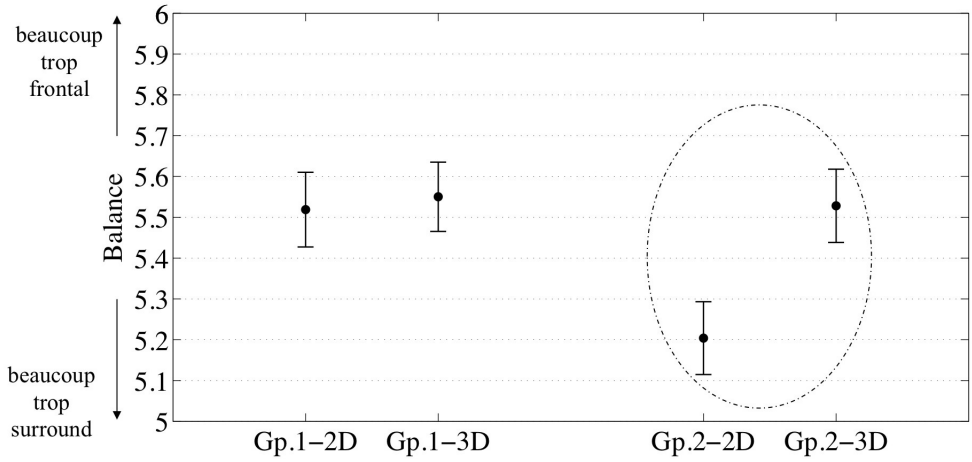


FIGURE 4.12 – Effet du Mode Visuel (2D vs. 3D-s) sur la perception de la balance frontal/surround en fonction du Groupe lors de l'expérience II. Moyennes et intervalles de confiance à 95%.

Influence de la Balance : mixage A vs. mixage B

Les sujets ont globalement bien perçu la différence entre les balances (B : $F(1, 42) = 156.687$, $p < 0.001$) avec une moyenne de 5.855 pour les mixages A et une moyenne de 5.046 pour les mixages B. Cependant, l'interaction Mode Visuel - Balance (V*B : $F(1, 42) = 0.445$, $p = 0.509$) n'est pas significative, ce qui veut dire que l'effet global du Mode Visuel a été le même indépendamment de la balance frontal/surround proposée aux sujets par les

expérimentateurs.

Etude du groupe 2, pour les séquences 1 et 2

Les Fig. 4.13 et 4.14 se concentrent sur les conditions où des différences significatives entre versions 2D et 3D-s ont pu être observées : les séquences 1 et 2, pour le groupe 2 situé au « Sweet Spot ».

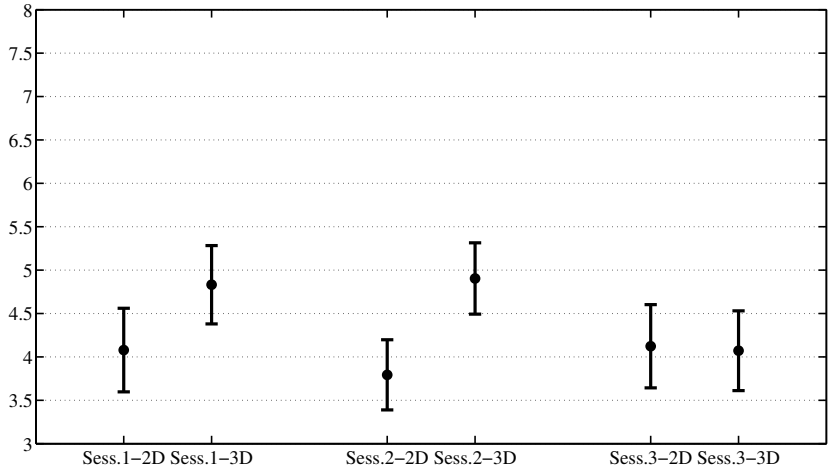


FIGURE 4.13 – Effet au « Sweet Spot » du Mode Visuel (2D vs. 3D-s) sur la perception de la balance frontal/surround en fonction de la session pour la Séquence 1 lors de l'expérience II. Moyennes et intervalles de confiance à 95%.

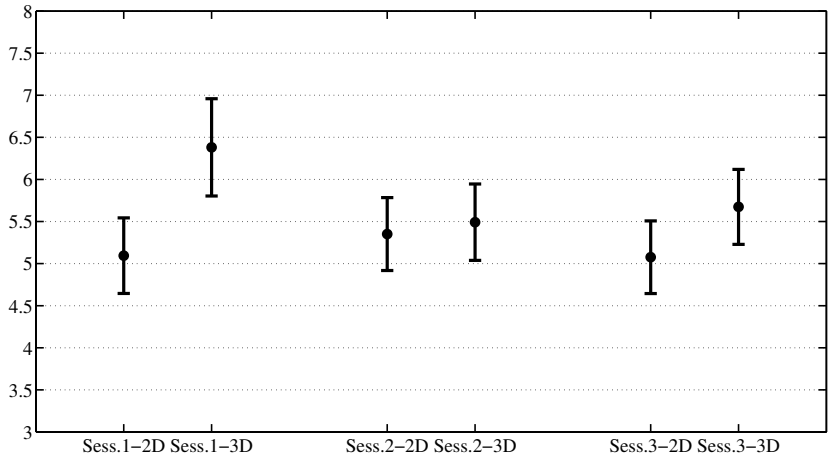


FIGURE 4.14 – Effet au « Sweet Spot » du Mode Visuel (2D vs. 3D-s) sur la perception de la balance frontal/surround en fonction de la session pour la Séquence 2 lors de l'expérience II. Moyennes et intervalles de confiance à 95%.

Nous constatons que :

— lors de la première session, l'effet de la stéréoscopie concerne les deux séquences, avec des différences moyennes sur l'échelle de notation de 0.75 et 1.29 pour les séquences 1 et 2 ;

— lors de la deuxième session, l'effet de la stéréoscopie ne concerne plus que la première séquence, avec une différence moyenne de 1.11 sur l'échelle de notation ;
— lors de la troisième session, l'effet de la stéréoscopie ne concerne cette fois-ci plus que la deuxième séquence, avec une différence moyenne de 0.60 sur l'échelle de notation. Cette différence n'a cependant pas été suffisamment importante pour que l'effet global de la stéréoscopie soit significatif dans la session 3 (cf. Fig. 4.11).

Ainsi, les différences concernent de moins en moins de séquences (2 séquences dans la session 1, 1 séquence dans la session 2 puis 1 séquence dans la session 3) et leur magnitude décroit d'une session à l'autre. Ces résultats sont cohérents avec ceux présentés dans la Fig. 4.11, dans laquelle on peut voir que l'effet de la stéréoscopie s'amenuise avec le temps.

4.4 Expérience III : Recherche de corrélations entre différences visuelles perçues et différences de balances perçues

L'expérience II a montré que la stéréoscopie n'avait d'effet sur la perception de la balance frontal/surround que pour 2 séquences sur 8. Une troisième expérience fut conduite pour déterminer si ces deux séquences étaient celles dont les différences entre versions 2D et 3D-s étaient les plus importantes en termes de perception visuelle.

4.4.1 Matériel et méthode

Le test s'est déroulé dans le même cinéma que pour l'expérience II. Les séquences étaient présentées par paire (d'abord la version 2D d'une séquence, puis sa version 3D-s) et sans son (pour éviter une influence du mixage proposé par les expérimentateurs).

Pour chaque présentation, les sujets devaient choisir sur une échelle une gradation parmi les 11 proposées en l'entourant. Les extrémités étaient labellisées « très semblable » et « très dissemblable » (respectivement égales à 0 et 10 pour l'analyse statistique). Les labels ne faisaient volontairement pas référence à la profondeur visuelle pour ne pas influencer les sujets. De plus, les sujets n'auraient pas forcément compris l'échelle de la même façon (certains sujets auraient peut-être plutôt associé la notion de profondeur visuelle à celle de profondeur de champ par exemple).

Chaque présentation durait 40 secondes (20 secondes pour la version 2D, 20 secondes pour la version 3D-s). A la fin d'une présentation, des lumières tamisées s'allumaient pour permettre aux sujets de reporter leur réponse sur la feuille. Au bout de 20 secondes, les lumières s'éteignaient et le stimulus suivant était présenté. Les sujets étaient priés de garder leurs lunettes 3D-s pendant la totalité de l'expérience et ils ont tous passé le test en même temps. Le test a été répété deux fois, sans pause entre les deux sessions, et a duré 16 minutes.

60 personnes (15 femmes et 45 hommes, âgés de 20 à 25 ans) ont pris part à l'expérience.

Il s'agissait d'étudiants en formation aux métiers du son et de l'image (Master « Image & Son » de l'Université de Brest). Comme dans l'expérience II, les sujets ont été divisés en 2 groupes : un proche de l'écran et l'autre au « Sweet Spot ».

4.4.2 Résultats

Les données brutes ont subi une transformée en z, afin de minimiser les différences inter-sujets d'utilisation de l'échelle (ITU-R BS.1116-1, 1994; ITU-R BS.1286, 1997).

Vérification des hypothèses pour une ANOVA

Un test de Kolmogorov-Smirnov a été effectué sur chacune des cellules (niveau de significativité à 5%) pour vérifier l'hypothèse de normalité. Le test a rejeté l'hypothèse d'une distribution normale pour seulement 7 cellules sur 32. Un test de Levene a également montré que la variance des cellules variait significativement d'un groupe à l'autre pour une seule combinaison de variables intra-sujets sur 16. De plus, le nombre d'observations par cellule est encore plus grand que dans l'expérience II (30 observations par cellule) et reste égal dans chaque cellule. L'utilisation d'une ANOVA est donc légitime. Un test de Mauchly a montré que l'hypothèse de sphéricité était violée pour le facteur « Séquence ». La valeur F de ce facteur a donc été corrigée, en utilisant la correction de Huynh-Feldt (Girden, 1991).

ANOVA

Les résultats de l'analyse de variance sont présentés dans le Tableau 4.4, avec les effets des 3 facteurs suivants :

Intra-sujets

- S : Séquence (8 niveaux) ;
- R : Répétition (2 niveaux).

Inter-sujets

- G : Groupe (2 niveaux : groupe 1 « Proche de l'écran » vs. groupe 2 « au Sweet Spot »).

Tableau 4.4 – Résultats de l'ANOVA pour l'expérience III.

Source	SS	DF	MS	F	Sig. p
S	2353.789	6.399	367.843	74.225	< 0.001
R	108.591	1	108.591	15.897	< 0.001
G	31.176	1	31.176	6.36207E+16	< 0.001
S*R	26.746	7	3.821	1.702	0.107
S*G	36.915	7	5.274	1.164	0.322
R*G	0.94	1	0.94	0.138	0.712
S*R*G	20.586	7	2.941	1.31	0.244

Les résultats montrent que tous les facteurs simples sont significatifs. Par contre, il n'y a aucune interaction significative.

Influence de la Séquence

L'effet de la Séquence est significatif (S : F(6.399, 371.136) = 74.225, p < 0.001). Un *post-hoc* LSD de Tukey montre que toutes les séquences sont significativement différentes entre elles, sauf les séquences 4, 6 et 8.

Les notes moyennes avec leur intervalles de confiance à 95%, obtenues pour l'évaluation des différences perçues de profondeur visuelle entre version 2D et version 3D-s séquence par séquence, sont présentées dans la Fig. 4.15.

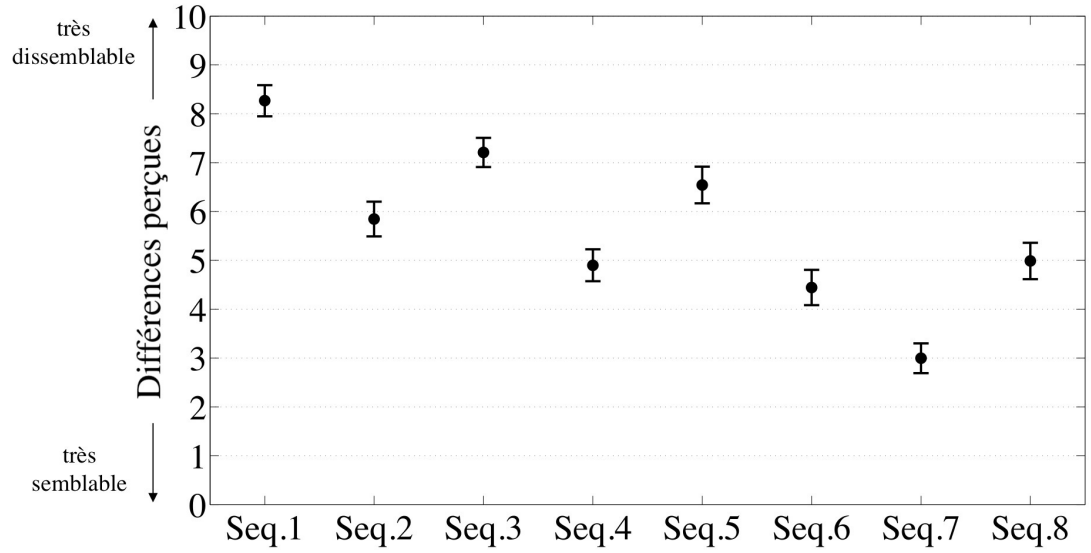


FIGURE 4.15 – Différences de profondeur visuelle perçue entre versions 2D et 3D-s pour chaque séquence lors de l'expérience III. Moyennes et intervalles de confiance à 95%.

Influence du Groupe : Proximité de l'écran vs. « Sweet Spot »

Les résultats montrent une influence significative du groupe (G : F(1, 58) = 6.36207E+16, p < 0.001), avec des moyennes égales à 5.829 pour le groupe 1 et 5.469 pour le groupe 2. Le groupe en proximité de l'écran a donc été plus sensible à la stéréoscopie que le groupe au « Sweet Spot », ce qui n'est pas surprenant puisque le relief est de plus en plus compressé lorsqu'on s'éloigne de l'écran (Moulin, 2015).

Influence de la Répétition

Les résultats montrent une influence significative de la répétition (R : F(1, 58) = 15.897, p < 0.001), avec des moyennes égales à 5.985 pour la première session et 5.313 pour la seconde session. Les sujets ont donc été plus sensibles à la stéréoscopie durant la première session.

La non-significativité de l'interaction "Séquence × Répétition" montre que l'effet de la répétition a été le même qu'importe la séquence. En effet, si nous ne retenons que les résultats de la seconde session, nous constatons que l'ordre des séquences en termes de différences perçues reste inchangé (voir Fig. 4.16)

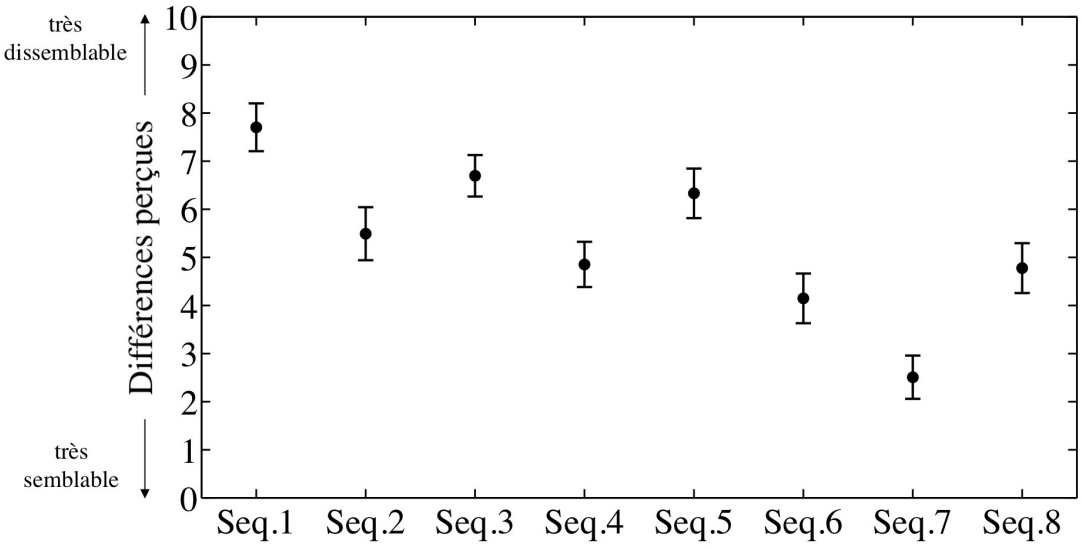


FIGURE 4.16 – Différences de profondeur visuelle perçue entre versions 2D et 3D-s pour chaque séquence lors de la seconde session de l'expérience III. Moyennes et intervalles de confiance à 95%.

La non-significativité de l'interaction "Groupe × Répétition" montre que l'effet de la répétition a également été le même pour les deux groupes (voir Fig. 4.17).

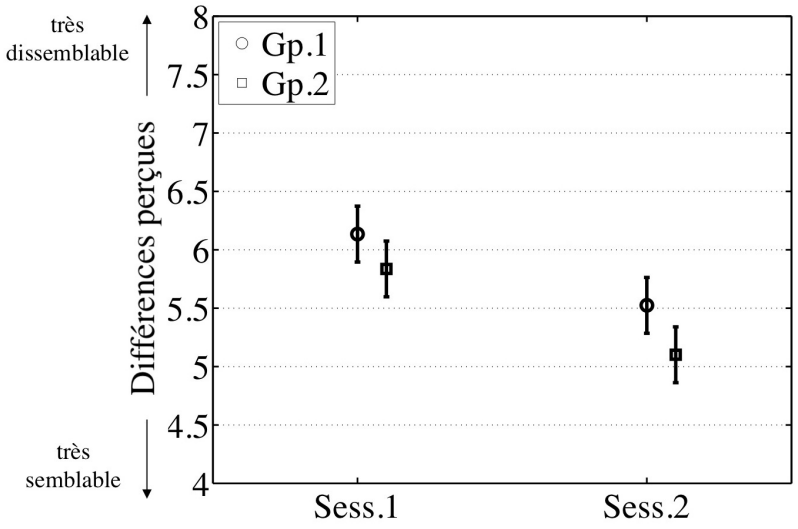


FIGURE 4.17 – Différences de profondeur visuelle perçue entre versions 2D et 3D-s pour chaque groupe dans les sessions 1 et 2 de l'expérience III. Moyennes et intervalles de confiance à 95%.

4.4.3 Recherche de corrélations

Corrélations entre l'expérience II et l'expérience III

Nous avions posé l'hypothèse que les deux séquences significatives de l'expérience II seraient sûrement celles ayant les différences visuelles les plus marquées entre versions 2D et 3D-s. Les résultats de l'expérience III montrent que la séquence 1 a en effet le plus haut degré de différences perçues. Cependant, la séquence 2 n'arrive que quatrième, après les séquences 3 et 5. D'ailleurs, le coefficient de corrélation rho de Spearman confirme qu'il n'y a aucune corrélation entre différences visuelles perçues et différences de balances perçues entre mixages 3D-s et mixages 2D ($\rho = 0.310$, p = 0.456 >> 0.05).

Dans l'expérience II, un effet global de la stéréoscopie n'avait été observé que pour le groupe 2 au « Sweet Spot » et nous pouvons donc regarder les corrélations entre expériences II et III uniquement pour ce groupe : une corrélation peut alors être observée, mais elle reste modeste ($\rho = 0.69$, p = 0.029).

Corrélations entre les tailles de boîte scénique et les résultats des expériences II et III

Le tableau 4.5 montre les tailles des boîtes scéniques mesurées pour les versions 3D-s de chaque séquence. Ces tailles sont déterminées en mesurant en pixels la différence entre la disparité de l'objet le plus loin et de l'objet le plus proche. Ces mesures permettent de décrire objectivement la « quantité » de stéréoscopie dans une séquence. Nous avions quand même tenu à mesurer cette « quantité » de stéréoscopie subjectivement avec l'expérience III, car Moulin (2015) avait montré que la boîte scénique ne permettait pas toujours de prédire correctement le degré de profondeur visuelle perçue par les sujets.

Séquence	boîte scénique (px)
1	31
2	24
3	31
4	12
5	24
6	18
7	8
8	12

Tableau 4.5 – Boîte scénique pour chaque séquence de l'expérience II

La corrélation avec l'expérience III est excellente ($\rho = 0.909$, p = 0.001), ce qui montre que la boîte scénique est dans le cadre de notre expérience un bon prédicateur de la sensation de différence visuelle perçue. Il n'y a cependant toujours pas de corrélation entre les valeurs de boîte scénique de chaque séquence et les différences globales de balances perçues entre mixages

3D-s et mixages 2D lors de l'expérience II (ρ = 0.242, p = 0.281). La corrélation devient cependant bien meilleure si on se limite aux différences de balances perçues au « Sweet Spot » (ρ = 0.85, p = 0.004). Ainsi, au « Sweet Spot », les différences de balances frontal/surround perçues entre versions 2D et 3D-s ont été d'autant plus importantes que les tailles des boîtes scéniques étaient grandes.

La Fig. 4.18 montre les différences de balances frontal/surround perçues entre versions 2D et 3D-s, pour chaque séquence et au « Sweet Spot », mesurées lors de l'expérience II en fonction de la taille de la boîte scénique. Nous remarquons que la différence de balances perçue est plus importante que ce que la taille de la boîte scénique laisserait présager pour la séquence 2 (entourée en rouge). Il est possible que les sujets aient eu un désir d'enveloppement lors de la version 3D-s plus prononcée pour cette séquence du fait qu'elle se déroule sur un bateau en pleine mer, un milieu par nature très enveloppant.

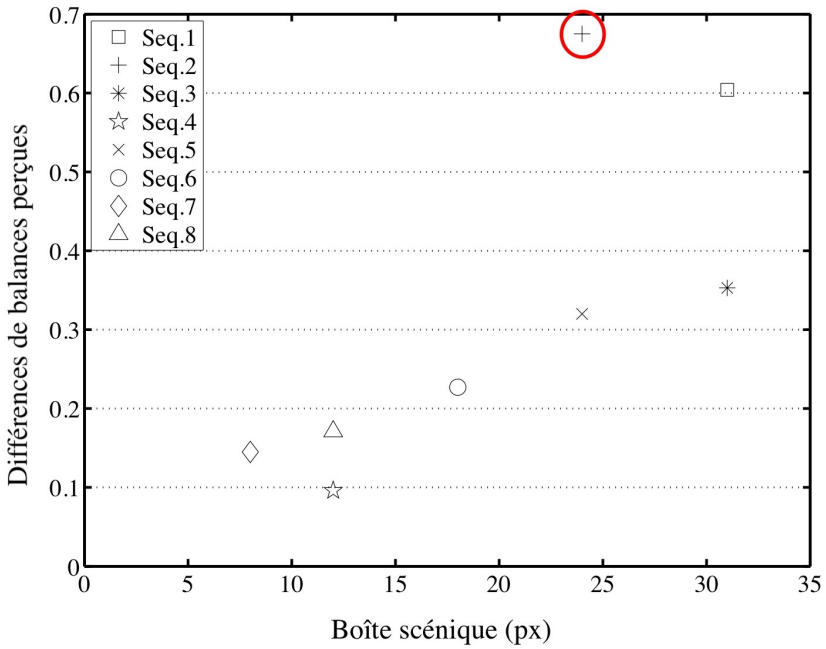


FIGURE 4.18 – Différences de profondeur visuelle perçue entre versions 2D et 3D-s mesurées pour chaque séquence au « Sweet Spot » en fonction de la taille de la boîte scénique (en pixels).

4.5 Discussion

4.5.1 Effet du Mode Visuel et dépendance à la Séquence et à la position dans la salle dans l'expérience II

Les résultats de l'expérience II montrent qu'il y a eu un effet global du Mode Visuel significatif : avec des images projetées en 3D-s, les mixages ont eu tendance à sonner plus frontaux, ce qui est cohérent avec plusieurs témoignages d'ingénieurs du son affirmant produire des

mixages plus « surround » pour la version 3D-s des films que pour leur version 2D (Gambier, 2010). Cependant, l'analyse des interactions a montré que le Mode Visuel n'était significatif que pour 2 séquences sur 8 et que pour le groupe 2 placé au « Sweet Spot ». Pour ce groupe, les différences entre mixages 3D-s et mixages 2D étaient respectivement égales à 0.6 et 0.7 pour les séquences 1 et 2, soit 6 et 7% de l'échelle. Ces différences, même si elles paraissent faibles, ne sont probablement pas anodines : par exemple, les sujets au « Sweet Spot » n'ont également utilisé que 6% de l'échelle pour différencier les mixages A et B de la séquence 1, quand bien même les expérimentateurs avaient pour le mixage A augmenté le gain des enceintes frontales de 2 dB et baissé le gain des enceintes surround de 5.2 dB par rapport au mixage B.

Le groupe 1 en proximité de l'écran se trouvait proche des enceintes frontales. Il est donc possible que la non-significativité du Mode Visuel pour le groupe 1 soit due au fait que les mixages, que l'image soit en 2D ou 3D-s, étaient toujours trop frontaux dans l'absolu. Les sujets n'auraient cependant pas utilisé l'extrémité haute de l'échelle tout simplement parce que même le plus frontal des mixages (un mixage 2.0 par exemple) ne peut pas sonner « beaucoup trop frontal », tant nous sommes habitués avec la stéréophonie 2.0 à des reproductions sonores uniquement frontales.

4.5.2 Dépendance au temps dans l'expérience II

Dans l'expérience II, l'effet de la stéréoscopie s'est atténué au cours du temps jusqu'à ne plus être globalement significatif lors de la dernière session. Plusieurs hypothèses peuvent expliquer cette tendance :

— l'évolution des sessions peut traduire un effet d'apprentissage du protocole (où les sujets apprennent à utiliser l'échelle de notation). Dans ce cas, il est préférable de ne garder que la dernière session du test : il n'y a alors aucun effet global de la stéréoscopie, même si un léger effet peut être observé pour la séquence 2 au « Sweet Spot ».

— comme les 44 sujets ont passé le test en même temps, l'ordre des stimuli était exactement le même pour tous. Pour minimiser cet effet d'ordre, l'expérience a été répétée trois fois, avec un ordre des stimuli aléatoire et différent d'une session à l'autre. La différence entre les sessions traduit donc peut-être l'effet d'ordre des stimuli. Dans ce cas, il est préférable de conserver les trois sessions : un effet de la stéréoscopie n'est alors observé que pour les séquences 1 et 2 au « Sweet Spot ».

— d'après les témoignages des sujets, l'atténuation de l'effet de la stéréoscopie au cours des trois sessions pourrait être plutôt due à l'accoutumance à la stéréoscopie et à la fatigue. En effet, certains sujets rapportent qu'ils ont été au cours du test de plus en plus habitués à la 3D-s, jusqu'à ne plus sentir de différences importantes entre les deux modes visuels. D'ailleurs, l'interaction Mode Visuel $*$ Répétition montre que

les impressions des sujets sur les mixages 2D sont restés constantes alors que leurs impressions sur les mixages 3D-s ont évolué jusqu'à devenir les mêmes que pour les mixages 2D dans la troisième et dernière session, comme si leur perception des images 3D-s devenait de plus en plus semblable à leur perception des images 2D. L'expérience III conforte cette explication, puisque les sujets ont été moins sensibles à la stéréoscopie pendant la seconde session que pendant la première session. La stéréoscopie ainsi que la monotonie engendrée par le visionnage répété 12 fois de chaque séquence lors de l'expérience II a pu également altérer les facultés analytiques des sujets et accélérer le processus d'adaptation. Dans ce cas, il est préférable de ne garder que les deux premières sessions du test : un effet global de la stéréoscopie n'est alors observé que pour les séquences 1 et 2 au « Sweet Spot ».

Nous constatons que l'effet de la stéréoscopie reste faible, qu'importe l'hypothèse retenue pour expliquer l'influence de la session.

4.5.3 Différences entre les expériences I et II

Comme l'expérience II, l'expérience I suggère que l'influence de la stéréoscopie est faible. En effet, seule 1 séquence sur 11 lors de la première session, puis 3 séquences lors de la seconde session, ont été mixées de façon différente selon que l'image était projetée en 2D ou en 3D-s.

Les résultats de l'expérience I présentent néanmoins plusieurs différences avec l'expérience II :

- la séquence 10, par exemple, a été mixée plus frontale en 3D-s qu'en 2D lors de la première session de l'expérience I, contrairement à notre sous-hypothèse 1. Cependant, la séquence n'était plus significative lors de la seconde session. Dans l'expérience II, la séquence a certes été à nouveau significative, mais cette fois-ci les sujets ont trouvé le mixage 3D-s « trop frontal », et ce à un degré supérieur que le mixage 2D correspondant, ce qui sous-entend qu'ils auraient préféré avoir plus de surround pour la version 3D-s de cette séquence que pour la version 2D.
- l'influence de la session a également été différente :
  - dans l'expérience I, il y avait plus de séquences significatives dans la seconde session que dans la première ;
  - dans l'expérience II, il y avait de moins en moins de séquences significatives au fur et à mesure des sessions.

Ces différences peuvent s'expliquer par le fait que la tâche était différente : les sujets était actifs lors de l'expérience I (dans la peau d'un mixeur, réglant eux-même la balance frontal/surround à l'aide d'un bouton rotatif) alors qu'ils étaient passifs dans l'expérience II (dans la peau de spectateurs, écoutant des balances déjà fixées). Les sujets n'ont donc pas forcément géré leur fatigue de la même façon. L'importante instabilité des sujets dans

l'expérience I d'une session à l'autre suggère également que la tâche était trop compliquée (ce qui expliquerait les incohérences avec l'expérience II) ou alors qu'elle aurait nécessité une phase d'apprentissage plus longue.

4.5.4 Aucune corrélation avec la profondeur visuelle perçue (expérience III), mais une bonne corrélation avec les tailles des boîtes scéniques des séquences

Aucune corrélation substantielle n'a pu être trouvée entre les différences de balances perçues dans l'expérience II et les différences visuelles perçues entre versions 2D et versions 3D-s de l'expérience III. Par contre, une forte corrélation a pu être observée entre les tailles des boîtes scéniques des séquences et les différences de balances perçues au « Sweet Spot » dans l'expérience II : plus la boîte scénique était grande, plus les différences de balances perçues entre versions 2D et 3D-s étaient importantes.

Il semble donc étrange que les résultats de l'expérience III n'aient pas présenté de meilleures corrélations avec ceux de l'expérience II, puisqu'il semble évident que les différences de balances perçues entre versions 2D et 3D-s d'une séquence sont bel et bien dépendantes de la « quantité » de stéréoscopie dans la version 3D-s de cette séquence. En focalisant dans l'expérience III l'attention des sujets non pas sur les différences globales perçues, mais sur les différences **de profondeur visuelle** perçues, peut-être aurions-nous observé une corrélation plus élevée avec les résultats de l'expérience II.

Aucune corrélation entre les tailles de boites scéniques et les mixages des sujets n'a pu être trouvée dans l'expérience I, ce qui semble supporter notre hypothèse que la tâche était trop compliquée.

4.6 Conclusion

En résumé, cette première série d'expériences montre que :

— les résultats confortent en partie la sous-hypothèse 1. Dans l'expérience 1, les sujets ont parfois produit des mixages plus « surround » lorsque l'image était projetée en s-3D que lorsqu'elle était projetée en 2D. L'expérience 2 a également montré que les mixages « sonnaient » globalement plus frontaux quand l'image était projetée en 3D-s que lorsqu'elle était projetée en 2D ;

— l'influence de la stéréoscopie est cependant limitée et n'apparaît que pour un petit nombre de séquences ;

— ce constat est valable, que le sujet soit dans la peau d'un mixeur ou dans la peau d'un spectateur ;

— cette influence s'amenuise avec le temps et ne concerne pas toutes les places dans la salle de cinéma ;

— La stéréoscopie modifie significativement la perception de la balance frontal/surround pour les séquences possédant de larges boîtes scéniques.

Chapitre 5

Influence de la stéréoscopie sur la perception des objets sonores

Sommaire

Ce chapitre se focalise sur la spatialisation des dialogues et effets sonores.

Dans l'expérience IV, nous avons souhaité poser la question de la pertinence d'une cohérence audiovisuelle dans le plan vertical et avons décidé d'étudier l'effet ventriloque en élévation, car très peu d'études ont été menées à ce sujet. Nous avons présenté à des sujets des séquences 3D-s montrant un homme en train de parler. Sa voix pouvait être reproduite sur différents haut-parleurs, qui créaient des disparités plus ou moins grandes en azimut et en élévation entre le son et l'image. Pour chaque présentation, les sujets devaient indiquer s'ils avaient perçu ou non la voix dans la même direction que la bouche de l'acteur. Les résul-

tats montrent que l'effet ventriloque est très efficace en élévation, et suggèrent qu'il n'est pas nécessaire de rechercher la cohérence audiovisuelle en élévation au cinéma.

Dans l'expérience V, nous avons tenté de vérifier la sous-hypothèse 2, à savoir que la stéréoscopie change nos attentes en termes de spatialisation des objets sonores (dialogues et effets), et qu'une plus grande cohérence spatiale entre le son et l'image est attendue lorsque l'image est projetée en 3D-s. Nous avons écarté la cohérence en élévation (au vu des résultats de l'expérience IV) et nous sommes focalisés sur l'azimut et la profondeur. Cette fois-ci, nous n'avons pas conduit d'étude classique sur l'effet ventriloque, car plusieurs études ont déjà été menées dans ces deux dimensions. Nous avons plutôt souhaité vérifier si la cohérence audiovisuelle en azimut et en profondeur pouvait améliorer la qualité d'expérience audiovisuelle des sujets, et si cette amélioration était différente selon que l'image était projetée en 2D ou en 3D-s. Cette expérience a fait l'objet d'une publication dans (Hendrickx *et al.*, 2015, accepté).

5.1 Expérience IV : Effet ventriloque avec des sources sonores variant à la fois en azimut et en élévation

5.1.1 Introduction

Dans le chapitre 3.1, nous avons vu que les nombreuses études sur l'effet ventriloque en azimut avaient permis de dégager plusieurs facteurs déterminant son efficacité :

- la disparité spatiale (l'effet décroît lorsque l'écart angulaire entre son et image augmente) ;
- la disparité temporelle (l'effet fonctionne mieux si le son et l'image sont synchrones) ;
- l'expérience du sujet (les sujets experts sont plus discriminants que les sujets naïfs) ;
- le « réalisme » de la combinaison son-image (l'effet fonctionne d'autant mieux que la combinaison son-image est réaliste et convaincante).

Il a également été montré que l'effet ventriloque dépendait de la précision spatiale du système auditif (l'effet fonctionne d'autant mieux que la précision de localisation est faible). Comme les performances de localisation sont moins bonnes dans le plan vertical que dans le plan horizontal, l'effet ventriloque devrait donc mieux fonctionner en élévation qu'en azimut.

Les études de Thurlow et Jack (1973) semblent conforter cette hypothèse : avec de larges disparités entre son et image dans le plan vertical, l'effet ventriloque fonctionnait bien mieux qu'avec de petites disparités dans le plan horizontal (voir chap. 3.1.4). Malheureusement, Thurlow n'a pas mesuré de « seuils à 50% » et ses résultats ne peuvent donc pas être comparés avec la littérature.

D'un autre côté, Werner *et al.* (2013) ont bel et bien mesuré des seuils à 50% dans le plan médian, mais leurs valeurs sont semblables à celles obtenues en azimut par d'autres études

($\approx 8-10°$). Werner conclut que la magnitude de l'effet ventriloque est similaire en élévation et en azimut. Nous pensons que cette similarité est plutôt due à la spécificité des conditions expérimentales de Werner, et qu'il est nécessaire de comparer des seuils mesurés en azimut et en élévation dans les mêmes conditions expérimentales pour pouvoir véritablement conclure.

Plusieurs études ont également suggéré que l'effet ventriloque fonctionnait mieux si le sujet prêtait moins attention à la position de la source sonore. Par exemple, lorsqu'une personne parle « dans la vraie vie », l'attention est plutôt focalisée sur le contenu sémantique du propos, et l'individu n'accorde vraisemblablement que très peu d'importance à la position spatiale de la voix.

Le but de cette expérience est de comparer la force de l'effet ventriloque en azimut et en élévation dans des conditions « réalistes ». Nous formulons l'hypothèse que :

— l'effet ventriloque est plus efficace en élévation qu'en azimut. Nous avons ainsi mesuré des seuils à 50% pour des stimuli sonores variant à la fois en azimut et en élévation, afin que l'effet ventriloque dans les deux dimensions puisse être comparé directement ;

— l'effet ventriloque peut fonctionner à des angles bien plus larges que ceux obtenus par Werner si les conditions expérimentales sont plus proches de la « vraie vie », avec :

- des combinaisons sons-images réalistes ;
- des sujets « naïfs » et non des experts entraînés pour la tâche ;
- l'attention du sujet focalisée sur le contenu sémantique des stimuli.

5.1.2 Matériel et méthode

Stimuli

Les séquences utilisées dans ce test montraient un jeune homme sur fond noir prononçant des phrases de 5 secondes (voir Fig. 5.1). Les séquences avaient été filmées en 3D-s à l'aide d'une caméra Panasonic AG-3DP1 et étaient projetées sur un écran face au sujet. Contrairement aux expériences I, II, III et V, les séquences étaient uniquement projetées dans leur version 3D-s (projeter les séquences à la fois en 2D et en 3D-s aurait rendu le test trop long).

Pour le contenu des phrases prononcées par l'acteur, nous nous sommes inspirés des corpus de stimuli habituellement utilisés dans les études sur le masquage (Martin *et al.*, 2012). Il s'agissait de phrases, en français, construites sur le modèle suivant : « Je m'appelle {nom}, ma couleur préférée est le {couleur} et j'habite à {ville} ». Il y avait trois noms possibles (Antoine, Clément, Pierre), trois couleurs possibles (rouge, vert, bleu) et trois villes possibles (Bordeaux, Lyon, Marseille). Avec toutes les combinaisons possibles de noms, couleurs et villes, le stimulus pouvait donc prendre $3 \times 3 \times 3 = 27$ formes différentes.

L'enregistrement des séquences a eu lieu dans une cabine de prise de son (acoustique mate) de l'Université de Brest, avec un microphone DPA 4006 placé 22 cm au-dessus de la bouche de l'acteur (pour que le microphone ne soit pas dans le champ de la caméra) et relié à une

FIGURE 5.1 – Capture d'écran d'une des séquences utilisées dans l'expérience IV. Toutes les séquences étaient semblables, seul le contenu des phrases prononcé par l'acteur changeait.

.

interface RME Fireface 800.

Système de reproduction

Le test subjectif s'est déroulé dans la salle 3D de l'université de Brest (acoustique mate). Les lumières avaient été éteintes pour minimiser l'influence d'éventuels indices visuels. Le sujet était assis au centre de la pièce.

Les images (25 i/s) étaient diffusées à l'aide d'un projecteur Epson EH-TW6000 sur un écran acoustiquement transparent, avec des lunettes 3D actives Epson ELPGS01. Le champ visuel des images projetées était de 33°.

La diffusion des stimuli et l'enregistrement des réponses des sujets étaient assurés par un logiciel programmé sous Max/MSP sur un ordinateur MacBook Pro relié à une interface RME MADIface USB.

Le système de diffusion sonore était composé de 28 enceintes Amadeus PMX4, alimentées par un convertisseur numérique-analogique D.O.Tec Andiamo 2.DA et des amplificateurs Audac DPA154. Chaque haut-parleur avait été filtré numériquement pour égaliser les réponses en fréquence. Pour chaque présentation, la voix de l'acteur était reproduite aléatoirement sur l'un des 28 haut-parleurs à un niveau sonore d'environ 65 dBA. La fréquence d'échantillonnage des signaux audio était de 48 kHz et la quantification était effectuée sur 24 bits.

Placement des haut-parleurs

Plusieurs études ont obtenu une symétrie gauche-droite pour l'effet ventriloque (Werner *et al.*, 2013; Wallace *et al.*, 2004) et il a donc été décidé de placer toutes les enceintes sur la droite du stimulus visuel.

Par contre, les résultats de Werner *et al.* (2013) suggèrent que les seuils à 50% dans le plan médian ne sont pas les mêmes selon que l'écart angulaire entre son et image est positif (son au-dessus de l'image) ou négatif (son en-dessous de l'image). Cependant, pour que la durée du test reste raisonnable, nous avons décidé d'étudier uniquement des écarts angulaires positifs (son au-dessus de l'image).

Un système de coordonnées sphériques bi-dimensionnel à deux pôles (avec l'azimut et l'élévation représentés par θ et ϕ respectivement) a été utilisé pour décrire les positions des haut-parleurs et des « seuils à 50% » sur une sphère de diamètre 2,40 m centrée sur la tête du sujet. Il est à noter que ce système de coordonnées diffère dans sa définition de l'azimut du système de coordonnées sphériques à un seul pôle plus fréquemment employé (voir Knudsen et Konishi (1979) ou Middlebrooks *et al.* (1989) pour une comparaison des deux systèmes).

Le stimulus visuel était projeté sur un écran droit devant le sujet, avec la bouche de l'acteur positionnée à azimut 0°, élévation 0° et 2,40 m de distance.

Le stimulus sonore pouvait être plus ou moins décalé par rapport au stimulus visuel le long de plusieurs arcs de cercles également centrés sur la tête du sujet. Les arcs de cercle pouvaient être plus ou moins inclinés par rapport au plan horizontal : l'angle que formaient au niveau de la bouche de l'acteur un arc de cercle et le plan horizontal est noté δ et appelé *orientation*. Pour que la durée du test reste raisonnable, 4 valeurs ont été retenues pour δ : 0° (décalage de la source sonore vers la droite), 45°, 67.5° (en diagonale) et 90° (vers le haut). Les 4 orientations δ sont représentées dans la Fig. 5.2.

Pour chaque orientation δ :

— une *indication de fusion* correspond à une situation où le sujet indique que la voix et la bouche de l'acteur lui semble provenir de la même direction ;

— l'angle au niveau de la tête du sujet entre le stimulus visuel (la bouche de l'acteur située droit devant le sujet) et le stimulus sonore (la voix de l'acteur) est appelé *écart angulaire* et noté Ψ ;

— la valeur de Ψ pour laquelle le pourcentage d'indications de fusion est égal à 50% (c'est-à-dire l'écart angulaire Ψ pour lequel la voix et la bouche semblent provenir de la même direction une fois sur deux) est appelé *seuil à 50%* et noté $\Psi_{50\%}$;

— La valeur de Ψ peut être décomposée en différences d'azimut et d'élévation entre le son et l'image : l'azimut et l'élévation correspondants sont respectivement notés θ et ϕ.

— Le seuil à 50% $\Psi_{50\%}$ peut être décomposé en différences d'azimut et d'élévation entre le son et l'image : l'azimut et l'élévation correspondants sont respectivement notés $\theta_{50\%}$ et $\phi_{50\%}$.

La Fig. 5.3 donne l'exemple d'une enceinte A positionnée le long de l'orientation δ 67.5° avec un écart angulaire $\Psi = 36°$.

Les coordonnées polaires peuvent être calculées à partir de δ et Ψ via la formule :

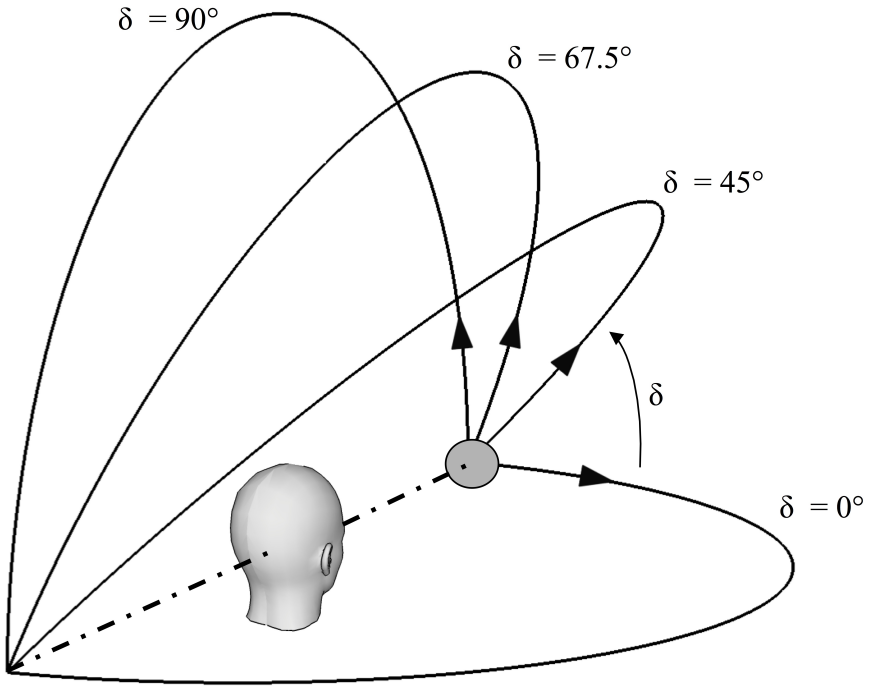


FIGURE 5.2 – Les 4 orientations δ, le long desquelles pouvait être décalé le stimulus sonore par rapport au stimulus visuel. Les 4 orientations étaient centrées sur la tête du sujet. Le stimulus visuel était toujours projeté à azimut 0°, élévation 0°, et est représenté sur la figure par un disque gris

.

$$azimut : \theta_A = \arcsin(\sin\Psi \times \cos\delta) = \arcsin(\sin 36^\circ \times \cos 67.5^\circ) \approx 13^\circ$$

$$Elevation : \phi_A = \arcsin(\sin\Psi \times \sin\delta) = \arcsin(\sin 36^\circ \times \sin 67.5^\circ) \approx 32.9^\circ$$

Le but de cette expérience est de définir pour chaque orientation δ le seuil à 50% $\Psi_{50\%}$. En plaçant des enceintes le long de chacune des 4 orientations δ (7 enceintes par orientation), 4 fonctions psychométriques peuvent être estimées, à partir desquelles les seuils à 50% et les pentes à 50% sont déterminés (la pente à 50% d'une fonction psychométrique correspond à la valeur de la pente au point où le pourcentage d'indications de fusion est égal à 50%).

Supposons que $\Psi_{50\%} = 32^\circ$ pour l'orientation δ 67.5°. $\Psi_{50\%}$ peut être décomposé en différences d'azimut et d'élévation :

$$\theta_{50\%} = \arcsin(\sin 32^\circ \times \cos 67.5^\circ) \approx 11.7^\circ$$

$$\phi_{50\%} = \arcsin(\sin 32^\circ \times \sin 67.5^\circ) \approx 29.3^\circ$$

Cela veut dire que si une enceinte est placée à azimut $\theta = 11.7^\circ$ et élévation $\phi = 29.3^\circ$, alors la voix de l'acteur sera perçue dans la même direction que sa bouche une fois sur deux lorsque sa voix sera diffusée sur cette enceinte.

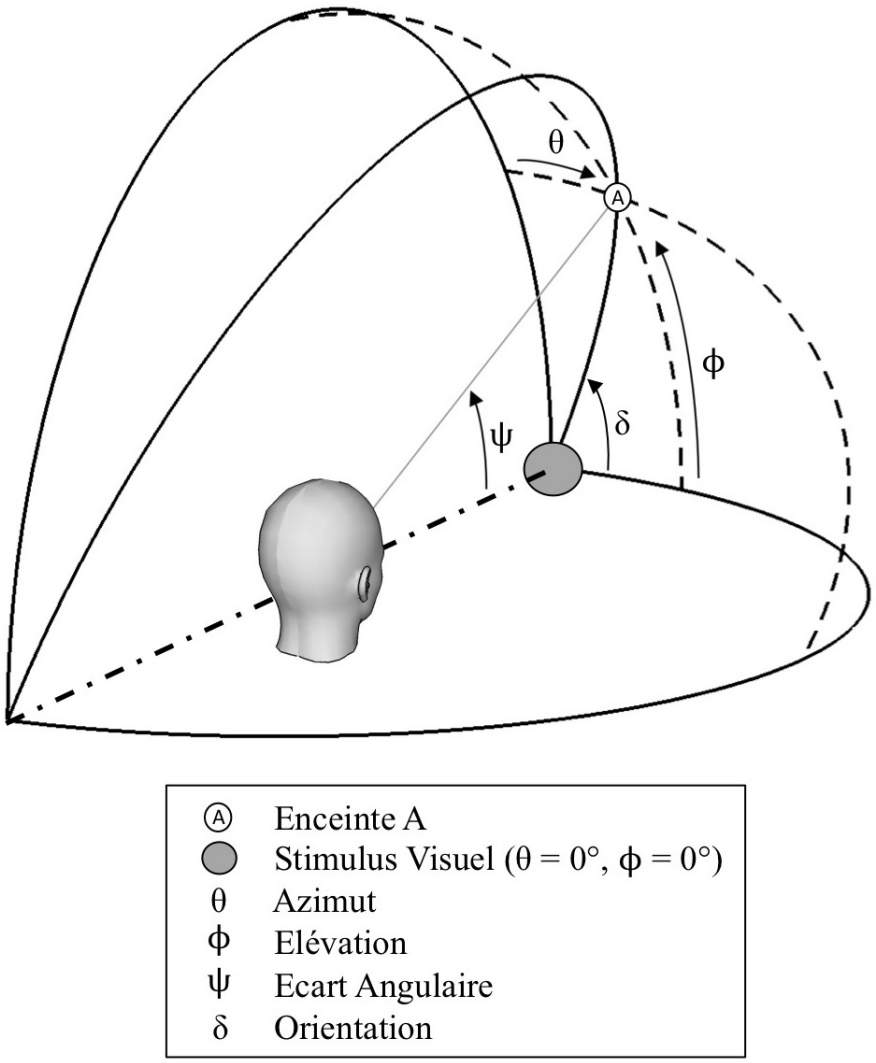


FIGURE 5.3 – Exemple d'une enceinte A d'orientation $\delta = 67.5°$ et d'écart angulaire $\Psi = 36°$.

Les valeurs d'orientation δ ont été déterminées à partir d'un test informel passé par les expérimentateurs, qui suggérait que le seuil à 50% $\Psi_{50\%}$ variait modérément entre $\delta = 0°$ et $\delta = 45°$ et substantiellement entre $\delta = 45°$ et $\delta = 90°$.

Il est à noter que le placement des enceintes à des angles d'orientation constants contraint l'azimut et l'élévation à varier d'une manière bien précise lorsque Ψ augmente le long d'une orientation δ :

- lorsque $\delta = 0°$, il n'y a pas de variation d'élévation au fur et à mesure que Ψ augmente (plan horizontal) ;
- lorsque $\delta = 45°$, azimut et élévation restent égaux au fur et à mesure que Ψ augmente ;
- lorsque $\delta = 67.5°$, l'élévation augmente plus rapidement que l'azimut au fur et à mesure que Ψ augmente ;
- lorsque $\delta = 90°$, il n'y a pas de variation d'azimut au fur et à mesure que Ψ augmente (plan médian) ;

Un test informel mené avec 6 sujets a permis d'estimer grossièrement la forme des fonc-

tions psychométriques. Le placement des enceintes a ainsi pu être optimisé en suivant les recommandations de Lam *et al.* (1999). La position exacte des enceintes est indiquée dans le Tableau 5.1. Il est à noter que la septième enceinte de l'orientation $\delta = 90°$ présentait un écart angulaire Ψ de 137° et se trouvait donc derrière le sujet.

	Orientation $\delta = 0°$			Orientation $\delta = 45°$			Orientation $\delta = 67.5°$			Orientation $\delta = 90°$		
Enceinte	θ	ϕ	Ψ	θ	ϕ	Ψ	θ	ϕ	Ψ	θ	ϕ	Ψ
1	5	0	5	7.1	7.1	10	4.2	10.2	11	0	10	10
2	10	0	10	10.5	10.5	15	7.5	18.4	20	0	19	19
3	14	0	14	13.3	13.3	19	10.7	26.6	29	0	27	27
4	18	0	18	15.4	15.4	22	13	32.9	36	0	34	34
5	23	0	23	17.4	17.4	25	16	41.7	46	0	43	43
6	27	0	27	20	20	29	18.5	50	56	0	90	90
7	31	0	31	24.6	24.6	36	22.5	67.5	90	0	137	137

TABLEAU 5.1 – Emplacements des enceintes, avec leur azimut θ, leur élévation ϕ et leur écart angulaire Ψ, pour chacune des 4 orientations δ.

À cause de l'écran et de la configuration de la pièce, les enceintes n'étaient pas toujours placées exactement à 2.40 m du sujet. L'erreur maximale de distance était de 16% (l'enceinte au-dessus des sujets, à élévation $\phi = 90°$). Cependant, les seuils de discrimination en distance obtenus par Ashmead *et al.* (1990) pour un stimulus à 2m (dont la distance variait mais à intensité constante) suggèrent que l'influence de ces erreurs sur les réponses des sujets devrait être négligeable.

Les enceintes n'étaient pas visibles par le sujet, sauf les enceintes 5, 6 et 7 de l'orientation 90° et de l'orientation 67.5°. Cependant, les lumières avaient été éteintes (y compris lorsque le sujet entrait dans la salle) pour minimiser l'influence visuelle. De plus, les séquences projetées n'étaient pas très lumineuses puisqu'il s'agissait d'un personnage sur fond noir. Il était également demandé au sujet de fixer uniquement la bouche du personnage à l'écran.

Pour les enceintes cachées derrière l'écran, il n'est pas impossible que l'écran ait provoqué un effet de diffusion spatiale, pouvant élargir la largeur apparente des sources sonores et "flouter" leur position spatiale. Il est donc possible que l'effet ventriloque ait été favorisé, et

donc surestimé pour ces conditions. Par contre, les effets de filtrage induits par l'écran ont été corrigés grâce à une égalisation.

Sujets et Protocole

8 sujets naïfs ont pris part à l'expérience (4 hommes, 4 femmes, âgés de 19 à 40 ans). Ils étaient rémunérés pour leur participation, et aucun d'entre eux n'avait déjà participé à un test d'écoute.

Un premier test (tâche A « sans question sémantique ») a été mené, dans lequel les sujets devaient répondre après chaque présentation à la question : « la voix et la bouche de l'acteur semblent-elles provenir de la même direction ? ». Une fois qu'ils avaient donné leur réponse, le stimulus suivant était automatiquement lancé.

Un test supplémentaire a été mené (tâche B « avec questions sémantiques »), dans lequel les sujets devaient après chaque présentation rapporter le nom, la couleur favorite et le lieu d'habitation du personnage avant de donner leur réponse sur la cohérence audiovisuelle. En cas de mauvaises réponses, l'essai en question était répété un peu plus tard. Cette tâche supplémentaire a été conduite afin de vérifier si le fait d'attirer l'attention du sujet sur le contenu sémantique (comme dans la « vraie vie ») pouvait permettre à l'effet ventriloque de mieux fonctionner.

En accord avec les recommandations de Lam *et al.* (1999), les sujets ont été interrogés 30 fois par enceinte pour chacune des deux tâches. L'ordre de diffusion des enceintes était aléatoire et différent pour chaque sujet. Pour chaque essai, un stimulus était choisi aléatoirement parmi les 27 combinaisons possibles de noms, couleurs et villes. Chaque tâche était divisée en deux sessions d'environ 1 heure, et tous les sujets ont passé les deux tâches sur 4 jours différents. Les sujets A, B, C et D ont commencé par la tâche A, tandis que les sujets E, F, G et H ont commencé par la tâche B.

5.1.3 Résultats

Pour estimer les fonctions psychométriques, une approche non-paramétrique basée sur un ajustement linéaire local (*local linear fitting* en anglais) a été utilisée. Cette méthode présente l'avantage de ne faire aucune hypothèse sur le véritable modèle décrivant les résultats (à part que la fonction psychométrique doit être continue et dérivable jusqu'à au moins l'ordre 2) tout en offrant des performances similaires voire meilleures que les modèles paramétriques courants (Zchaluk et Foster, 2009).

Comme le taux d'erreur était extrêmement bas (inférieur à 1% pour chaque sujet) durant la tâche B, il a été décidé d'ignorer les essais pour lesquels les sujets avaient donné de mauvaises réponses aux questions sémantiques.

Influence de l'orientation δ

La Fig. 5.4 montre les seuils à 50% $\Psi_{50\%}$ en fonction de l'orientation δ, pour chaque sujet, durant la tâche A (sans question sémantique). $\Psi_{50\%}$ ne pouvait pas toujours être déterminé à δ = 90° : pour les sujets B et F, le pourcentage d'indications de fusion était toujours supérieur à 50%, qu'importe l'enceinte (même pour l'enceinte derrière le sujet, à Ψ = 137°, le pourcentage d'indications de fusion était égal à 85% et 77% respectivement). La valeur 137° a été retenue pour la figure, mais il est probable que le pourcentage d'indications de fusion ait été supérieur à 50% à des valeurs d'écart angulaire plus grandes, peut-être même dans tout le plan médian.

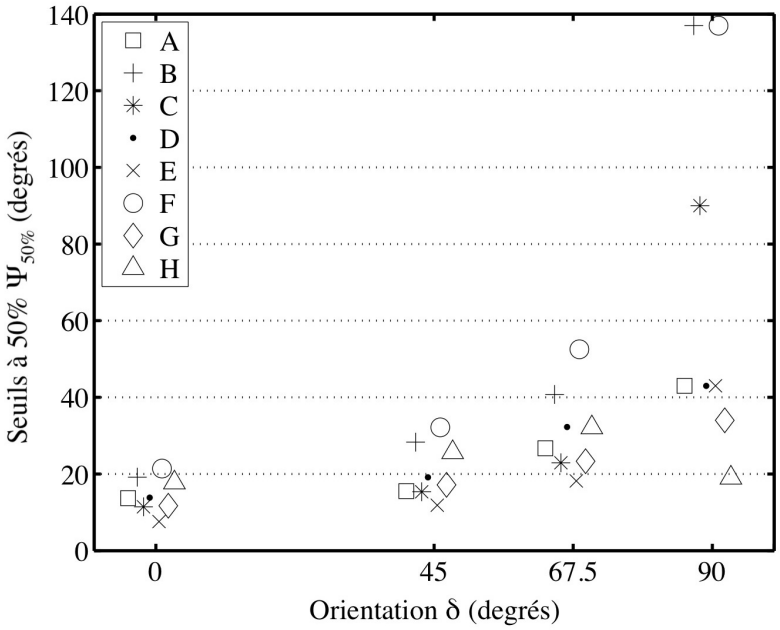


FIGURE 5.4 – Seuils à 50% $\Psi_{50\%}$ pour chaque sujet en fonction de l'orientation δ. Tâche A.

La Fig. 5.5 (résultats du sujet C à 90° d'orientation δ) montre un autre exemple où le seuil à 50% $\Psi_{50\%}$ ne pouvait pas être déterminé précisément : la pente de la fonction psychométrique étant pratiquement horizontale autour du point à 50% d'indications de fusion, l'imprécision sur l'estimation de $\Psi_{50\%}$ était considérable (l'intervalle de confiance allait de 40.6° à 107.3°).

Il a donc été décidé de ne pas utiliser les seuils à 50% pour les données obtenues à 90° d'orientation δ, mais plutôt d'indiquer la valeur maximale d'écart angulaire Ψ pour laquelle nous avions pu observer un pourcentage d'indications de fusion supérieur à 50%. Dans le cas du sujet C, la valeur Ψ = 90°, correspondant à la sixième enceinte de l'orientation δ 90°, a été indiquée. Comme les fonctions psychométriques étaient décroissantes et monotones pour tous les sujets, il est probable que les véritables seuils à 50% aient été en fait à des écarts angulaires plus importants que ceux indiqués.

Sur la Fig. 5.4, les seuils à 50% sont déjà relativement dispersés à 0° d'orientation δ, allant de 7° (sujet E) à 21° (sujet F) : certains sujets peuvent donc tolérer des écarts angulaires jusqu'à trois fois supérieurs que d'autres sujets.

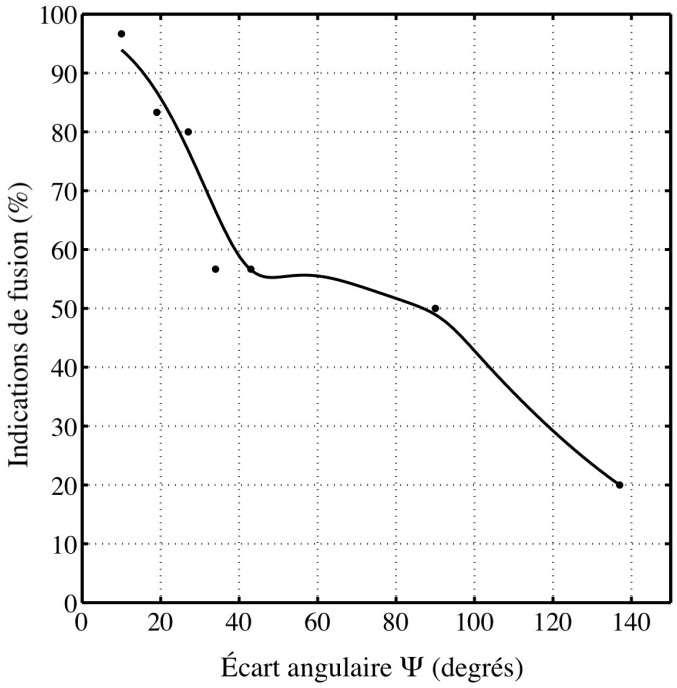


FIGURE 5.5 – Fonction psychométrique à 90° d'orientation δ pour le sujet C, durant la tâche A. Les points représentent les données obtenues pour chaque enceinte.

Lorsque l'orientation δ augmente de 0° à 45°, tous les sujets montrent à peu près la même tendance, avec une augmentation légère des seuils (moyenne = +6°).

Lorsque l'orientation δ augmente de 45° à $67,5°$, les seuils à 50% augmentent plus rapidement, et ce pour tous les sujets (moyenne = +10°, ce qui veut dire que l'augmentation moyenne du seuil est deux fois plus grande qu'entre $\delta = 0°$ et $\delta = 45°$, quand bien même la différence d'orientation δ est deux fois plus petite). Cependant, la vitesse d'augmentation du seuil varie significativement d'un sujet à l'autre : par exemple, elle est de +20° pour le sujet F, alors qu'elle n'est que de +6° pour le sujet G. Ces différentes vitesses accentuent la variabilité inter-sujet des seuils à 50% : à 67,5° d'orientation δ, deux sujets (B et F) présentent de bien plus grands seuils (41° et 53° respectivement) que les autres sujets (dont les seuils à 50% varient de 18° à 32°).

Lorsque l'orientation δ augmente de $67,5°$ à 90°, les seuils à 50% augmentent encore plus vite pour la plupart des sujets, avec des différences encore plus marquées entre les vitesses d'augmentation : +96° pour le sujet B, mais seulement +12° pour le sujet G. Cela entraine une dispersion importante des seuils à 50% à 90° d'orientation δ : de 19° pour le sujet H jusqu'à 137° pour les sujets B et F. Le sujet H est l'unique cas de figure où, étrangement, le seuil à 50% décroit de 9.5°. Il est à noter que les sujets A, D et E présentent le même seuil à 90° d'orientation δ, égal à 43°. Cependant, cette égalité est vraisemblablement due à

notre définition particulière du « seuil » pour cette orientation, et les véritables seuils à 50% sont probablement différents d'un sujet à l'autre et éparpillés entre 43° (cinquième enceinte de l'orientation) et 90° (sixième enceinte de l'orientation).

Une tendance similaire a été obtenue pour la tâche B avec questions sémantiques (cf. Fig. 5.6), bien que la différence de seuils à 50% entre $\delta = 67.5°$ et $\delta = 90°$ soit moins marquée. À nouveau, certains sujets (A, C, D et E) présentent le même seuil à 90° d'orientation δ, égal à 34°, mais les véritables seuils à 50% sont probablement différents d'un sujet à l'autre et éparpillés entre 34° (quatrième enceinte de l'orientation) et 43° (cinquième enceinte de l'orientation).

En résumé, les résultats des deux tâches A et B montrent que :

— les seuils à 50% augmentent strictement lorsque l'orientation δ augmente de 0° à 90° (sauf pour un sujet) ;

— cette augmentation est plus ou moins rapide selon le sujet, ce qui a pour effet d'élargir la variabilité inter-sujet au fur et à mesure que l'orientation δ augmente de 0° à 90°. A 0° d'orientation δ, les seuils à 50% sont déjà dispersés (de 7° à 21°), mais cette dispersion est finalement modérée par rapport à celle observée à 90° d'orientation δ, où les seuils à 50% vont de 19° jusqu'à 137°.

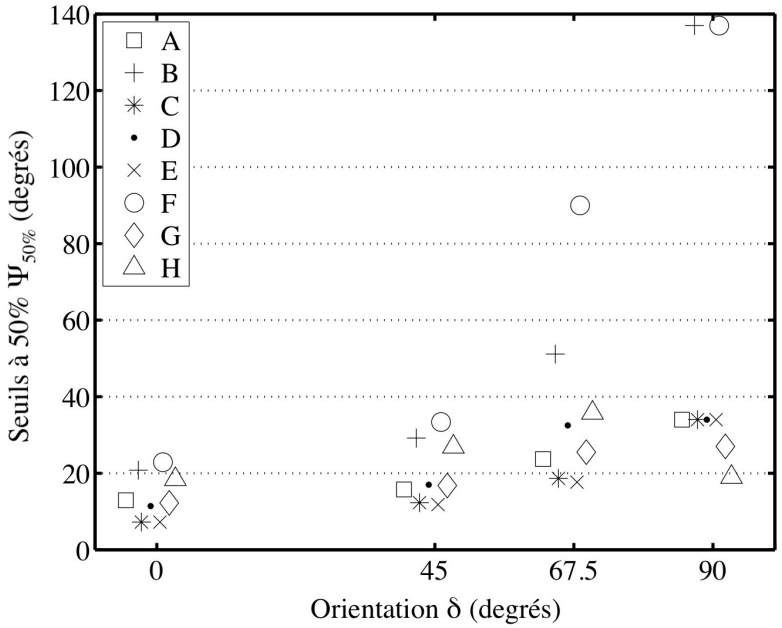


FIGURE 5.6 – Seuils à 50% $\Psi_{50\%}$ pour chaque sujet en fonction de l'orientation δ. Tâche B

Influence de la tâche : « sans question sémantique » (tâche A) vs. « avec questions sémantiques » (tâche B)

Des tests de Wilcoxon ont été appliqués aux données pour déterminer si la tâche avait eu un impact significatif sur les réponses des sujets :

— un premier test a comparé les seuils à 50% et les pentes obtenues aux orientations $\delta = \{0°, 45°, 67.5°\}$. L'influence de la tâche s'est révélée non significative ($p = 0.976$ pour les seuils à 50% et $p = 0.224$ pour les pentes).

— un second test de Wilcoxon a été mené sur les résultats obtenus à 90° d'orientation δ. Le test comparait les pourcentages d'indications de fusion obtenus pour chaque enceinte (cf. Fig. 5.7), en intégrant les résultats obtenus pour tous les sujets et toutes les enceintes de l'orientation. L'influence de la tâche s'est révélée significative ($p = 0.011$), avec plus d'indications de fusion lors de la tâche A « sans question sémantique » (64% en moyenne) que lors de la tâche B « avec questions sémantiques » (59% en moyenne).

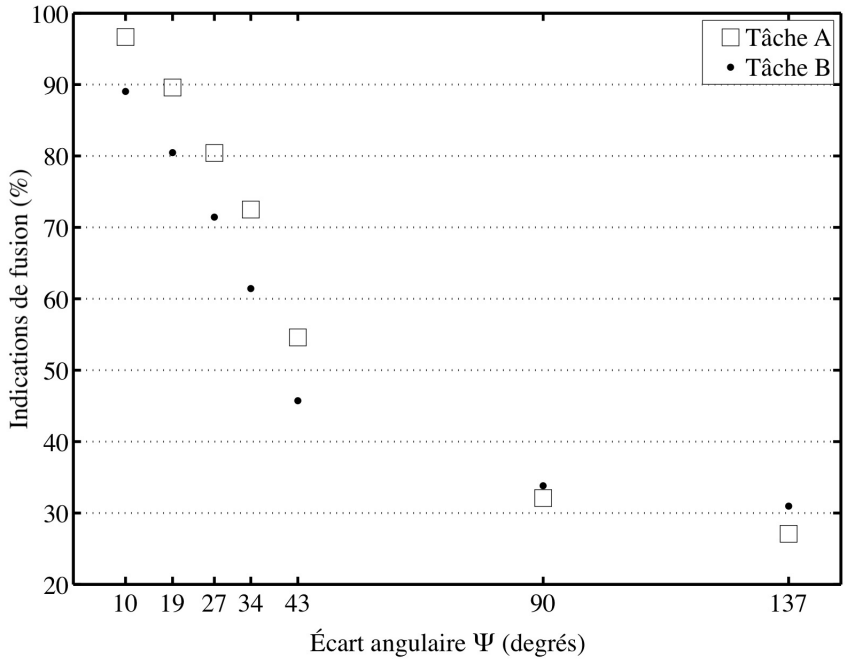


FIGURE 5.7 – Pourcentage d'indications de fusion pour chaque enceinte, à 90° d'orientation δ. Moyennes sur l'ensemble des sujets.

Influence de l'azimut et de l'élévation

La Fig. 5.8 montre les seuils à 50% $\Psi_{50\%}$ décomposés en azimut $\theta_{50\%}$ et en élévation $\phi_{50\%}$, pour la tâche A. L'aire sous chaque courbe correspond à la zone dans laquelle l'effet ventriloque a fonctionné plus d'une fois sur deux.

Au fur et à mesure que l'orientation δ augmente, les variations d'azimut sont modérées comparées aux variations d'élévation. Un test de Wilcoxon, intégrant les résultats des deux tâches A et B, montre même qu'il n'y a pas de différence significative d'azimut entre les orientations $\delta = 0°$ et $\delta = 45°$ ($p = 0.820$).

À une certaine orientation comprise entre $\delta = 45°$ et $\delta = 67.5°$, l'azimut du seuil à

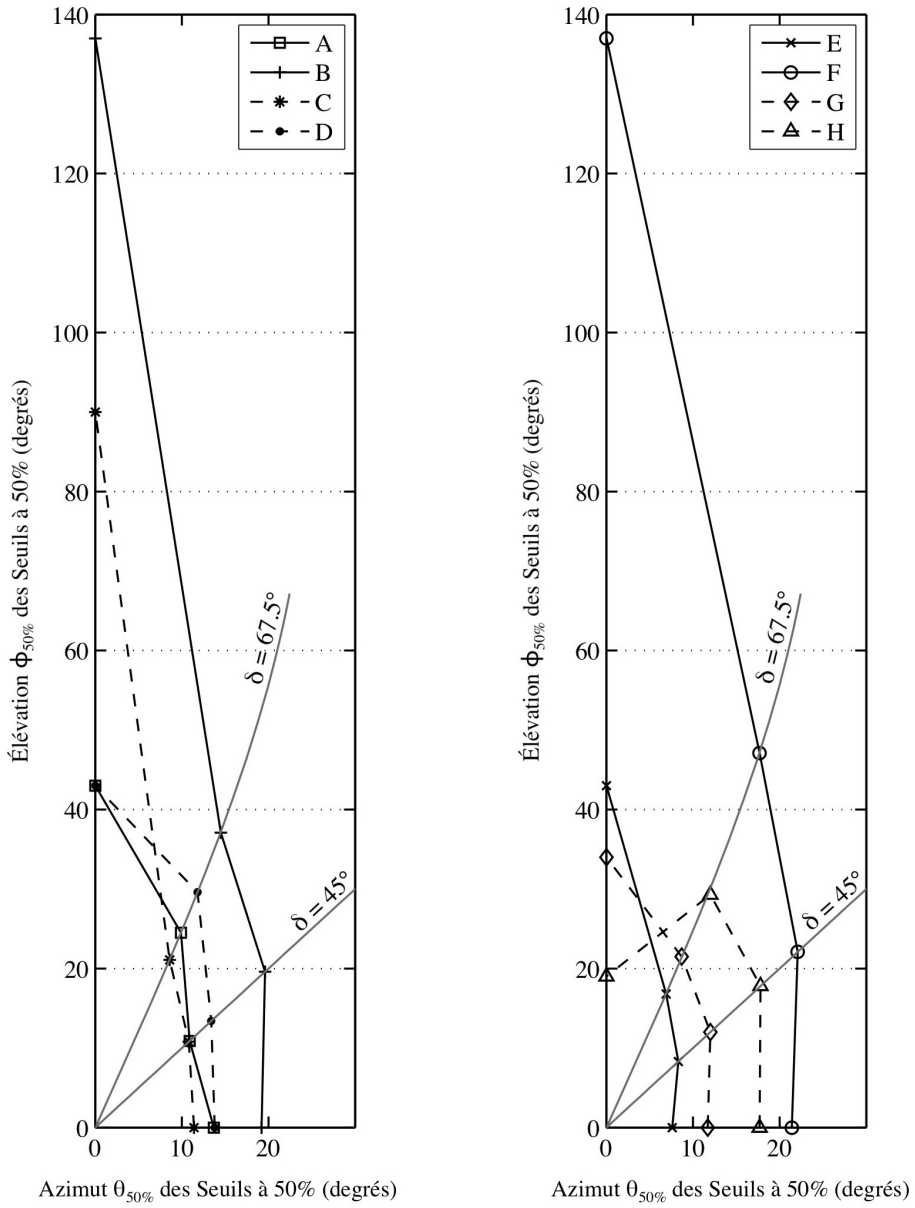


FIGURE 5.8 – Azimut $\theta_{50\%}$ et élévation $\phi_{50\%}$ des seuils à 50% pour les sujets A, B, C et D (figure de gauche) et les sujets E, F, G et H (figure de droite). Tâche A. Pour plus de lisibilité, les résultats ont été répartis sur deux diagrammes différents. L'axe des abscisses correspond à 0° d'orientation δ tandis que l'axe des ordonnées correspond à 90° d'orientation δ. L'aire sous chaque courbe correspond à la zone dans laquelle l'effet ventriloque a fonctionné plus d'une fois sur deux.

50% $\theta_{50\%}$ commence à décroître. Cependant, comme on peut le voir sur la Fig. 5.8, cette décroissance est faible comparée à l'augmentation simultanée de l'élévation $\phi_{50\%}$.

Ainsi, l'efficacité de l'effet ventriloque dépend uniquement de la différence d'azimut entre les stimuli visuel et sonore sur une large amplitude d'orientations (de $\delta = 0°$ à au moins $\delta = 45°$). Au-delà de 45°, la différence d'élévation doit également être prise en compte, mais les variations d'élévation ont bien moins d'impact sur l'effet ventriloque que les variations d'azimut.

5.1.4 Discussion

L'effet ventriloque fonctionne mieux en azimut qu'en élévation

Les résultats de la présente étude confortent l'hypothèse que l'effet ventriloque fonctionne mieux dans le plan vertical que dans le plan horizontal : les seuils à 50% sont en moyenne 4 fois plus grands à 90° d'orientation δ (plan vertical) qu'à 0° d'orientation δ (plan horizontal).

Ces résultats sont en accord avec les études précédentes de Thurlow et Jack (1973), ainsi qu'avec les performances de localisation du système auditif : comme la précision spatiale est moins bonne dans le plan vertical que dans le plan horizontal, l'influence de la position du stimulus sonore décroît et de plus larges écarts angulaires sont ainsi tolérés.

Cependant, une importante variabilité inter-sujet a pu être constatée, puisque les seuils à 50% pouvaient être de 1.1 jusqu'à 8 fois plus grands dans le plan vertical que dans le plan horizontal selon le sujet.

Pour des directions « obliques » (à 45° ou 67,5° d'orientation δ), les stimuli sonore et visuel présentaient à la fois des différences d'azimut et d'élévation. Cependant, les résultats montrent que les variations d'élévation avaient très peu d'impact sur l'effet ventriloque et que les seuils à 50% dépendaient principalement des différences d'azimut.

L'effet ventriloque peut fonctionner à des angles verticaux très larges

Nous avions également formulé l'hypothèse que l'effet ventriloque pouvait fonctionner à des angles bien plus importants que ceux obtenus par Werner *et al.* (2013) si les conditions expérimentales étaient plus proches de la « vraie vie ». Dans la présente expérience :

— les sujets étaient « naïfs », alors que ceux de Werner étaient expérimentés et entraînés pour la tâche à accomplir. Or, il a été montré que les seuils pouvaient être jusqu'à deux fois plus grands avec des sujets « naïfs » qu'avec des sujets « experts » ;

— la combinaison son-image des stimuli était plus réaliste que celle de Werner (qui utilisaient des salves de bruit blanc et des enregistrements de saxophone associés à des allumages/extinctions de LEDs blanches). Or, Thurlow et Jack (1973) ont montré sur une séquence de 5 minutes qu'une combinaison de stimuli réalistes pouvait multiplier par trois la durée moyenne durant laquelle l'effet ventriloque fonctionnait par rapport à une combinaison abstraite ;

— le son pouvait provenir de 28 directions différentes, alors que l'expérience de Werner n'utilisait que 4 enceintes différentes. Il a donc été probablement plus difficile pour les sujets d'« apprendre » la position des enceintes.

Les résultats montrent que les seuils à 50% dans le plan vertical étaient en effet bien plus grands (de 19° à des valeurs d'angles dépassant les 137°) que ceux obtenus par Werner *et al.* (2013) (8° et 10°). Pour deux sujets, il est même probable que l'effet ventriloque ait fonctionné dans la totalité du plan médian, puisque les pourcentages d'indications de fusion

étaient de 85% et 77% même quand le stimulus sonore était diffusé dans leur dos ($\Psi = 137°$ à 90° d'orientation δ).

Les fluctuations d'attention influencent l'effet ventriloque, mais uniquement dans le plan médian

Nous avions également formulé l'hypothèse que focaliser l'attention du sujet sur le contenu sémantique des stimuli permettrait à l'effet ventriloque de fonctionner à des écarts angulaires plus importants entre son et image. Cependant, la plupart des sujets ont rapporté que la mémorisation des noms, couleurs favorites et lieux d'habitations du personnage était une tâche simple qui n'avait pas détourné leur attention des disparités spatiales. D'autres sujets ont rapporté que les premiers mots « je m'appelle » (qui étaient les mêmes pour les 27 phrases) suffisaient à se faire une opinion sur la question « la voix et la bouche de l'acteur semblent-elles provenir de la même direction ? », et qu'ils pouvaient donc se focaliser sur le contenu sémantique pour la suite de la phrase. Les résultats ont montré qu'il n'y a effectivement pas eu d'influence significative de la tâche aux orientations $\delta = \{0°, 45°, 67.5°\}$.

Cependant, l'effet ventriloque a fonctionné significativement mieux durant la tâche A « sans question sémantique » que durant la tâche B « avec questions sémantiques » à 90° d'orientation δ. Ce phénomène ne peut résulter d'un effet d'apprentissage puisque l'ordre des tâches n'était pas le même pour tous les sujets. Forcer les sujets à se concentrer sur le contenu sémantique durant la tâche B a peut-être maintenu leur niveau de stimulation à un plus haut degré, les rendant ainsi plus discriminants sur la durée par rapport à la tâche A.

Même si ces résultats contredisent notre hypothèse de départ, ils montrent tout de même que des fluctuations d'attention du sujet peuvent avoir une influence significative sur l'effet ventriloque. A 0°, 45° et 67,5° d'orientation δ, ces fluctuations sont négligeables, mais ne le sont plus à 90° d'orientation δ.

Une pondération différente des facteurs déterminant l'effet ventriloque pourrait expliquer pourquoi la variabilité inter-sujet augmente au fur et à mesure que l'orientation δ augmente

Une hypothèse, très semblable au modèle proposé par Thurlow et Jack (1973), pourrait expliquer les tendances observées.

Le fait qu'un sujet « fusionne » un stimulus sonore avec un stimulus visuel est une décision complexe qui repose sur plusieurs facteurs tels que la position du stimulus sonore par rapport au stimulus visuel, à quel point le sujet estime que les deux stimuli « vont bien ensemble » (réalisme et crédibilité de la combinaison son-image), et à quel point le sujet prête attention à la position de la source sonore.

L'influence de ces facteurs est plus ou moins pondérée en fonction de la situation. Par

exemple, si la précision de localisation auditive est faible, le sujet hésitera parmi un nombre important de directions pour la localisation du son. Si la combinaison son-image est très convaincante, alors le sujet présumera que la direction la plus probable pour le stimulus sonore est celle de la source visuelle. Ainsi, l'influence du facteur « position de la source sonore » décroît au profit du facteur « réalisme de la combinaison son-image ».

Tandis que certains facteurs sont relativement constants d'un sujet à l'autre, d'autres facteurs présentent une grande variabilité inter-sujet :

— des études ont montré que les performances de localisation étaient comparables d'un sujet à l'autre (Makous et Middlebrooks, 1990) ;

— la « présomption d'unité » (i.e. à quel point un sujet estime qu'un son et un image vont « bien ensemble » et qui est en grande partie liée au réalisme de la combinaison son-image) dépend étroitement de l'expérience du sujet et de son passé avec des situations semblables (Warren *et al.*, 1981) ;

— l'attention prêtée aux informations du système auditif peut fortement varier d'un sujet à l'autre (Giard et Peronet, 1999).

Ainsi, si la pondération associée à un facteur hautement subjectif (tel que le réalisme de la combinaison son-image ou l'attention du sujet) augmente, alors nous pouvons formuler l'hypothèse que la variabilité inter-sujet augmentera également.

Dans le plan horizontal, les performances de localisation sont bonnes. Ainsi, tant qu'il y a un minimum de différences azimutales entre le son et l'image, l'effet ventriloque est fortement influencé par la position horizontale de la source sonore par rapport à la source visuelle. Ceci a plusieurs conséquences :

— l'effet ventriloque est limité ;

— une variation d'élévation de la source sonore n'a pas d'effet prononcé car le manque de précision en localisation verticale fait de cette variation un indice spatial négligeable par rapport à la différence d'azimut ;

— l'influence des autres facteurs tels que le réalisme de la combinaison son-image ou l'attention du sujet est réduite. Ainsi, la variabilité inter-sujet observée est modérée, et les fluctuations d'attention (tâche A vs. tâche B) sont négligeables ;

Cependant, au fur et à mesure que l'orientation δ augmente, il y a de plus en plus de différences d'élévations et de moins en moins de différences d'azimut entre les stimuli visuel et sonore. Comme la localisation est moins précise en élévation, l'influence de la position du stimulus sonore décroît, ce qui a deux conséquences :

— l'effet ventriloque fonctionne à des écarts angulaires plus larges ;

— l'influence des autres facteurs, tels que le réalisme de la combinaison son-image ou les fluctuations d'attention du sujet, augmente. Comme ces facteurs sont très subjectifs, une variabilité inter-sujet plus grande est observée. Une telle hypothèse expliquerait également pourquoi le facteur « attention » (tâche A vs. tâche B) n'est devenu signifi-

catif qu'à 90° d'orientation δ

La Fig. 5.9 montre les fonctions psychométriques obtenues pour le sujet D durant la tâche A. Les pentes déterminent à quel point l'efficacité de l'effet ventriloque varie avec l'écart angulaire Ψ. Une pente très raide signifie que le seuil à 50% sépare clairement l'écart angulaire Ψ en deux zones : une zone ou l'effet ventriloque marche en permanence (100% d'indications de fusion) et une zone où l'effet ne fonctionne jamais (0% d'indication de fusion). Si la pente est moins raide, cela signifie qu'il existe une zone d'« incertitude » autour du seuil à 50%, correspondant à un intervalle d'écarts angulaires pour lequel l'opinion du sujet varie d'une présentation à l'autre. Sur la Fig. 5.9, les pentes sont raides et semblables à 0° et 45° d'orientation δ. Cependant, au fur et à mesure que l'orientation δ augmente de 45° à 90°, la pente devient de moins en moins raide. Cette tendance, qui a été observée pour tous les sujets, conforte l'hypothèse que la décision des sujets est de plus en plus influencée par des facteurs fluctuants, tels que l'attention du sujet, au fur et à mesure que l'orientation δ augmente de 45° à 90°.

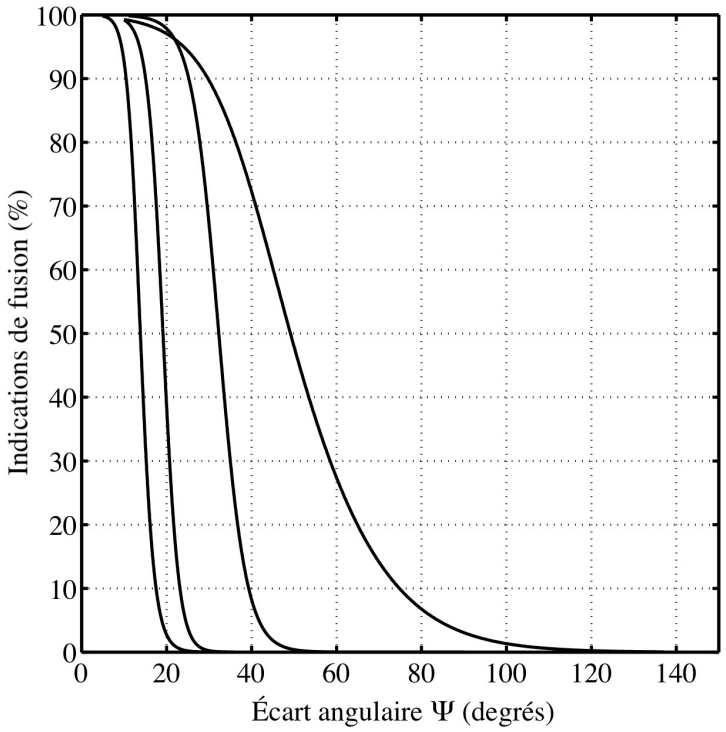


FIGURE 5.9 – Fonctions psychométriques obtenues pour le sujet D durant la tâche A pour chaque orientation δ. De gauche à droite : $\delta = 0°$, $\delta = 45°$, $\delta = 67.5°$, $\delta = 90°$.

5.1.5 Conclusion

Synthèse des résultats

Les résultats ont montré que l'effet ventriloque fonctionnait bien mieux en élévation qu'en azimut (en accord avec les performances de localisation du système auditif) et pouvait fonctionner à des écarts angulaires très élevés (certains sujets continuaient de percevoir la voix du personnage sur sa bouche même lorsque le son était diffusé dans leur dos). Cependant, la variabilité inter-sujet était plus grande en élévation qu'en azimut.

Dans une autre tâche, les sujets devaient répondre à des questions sur le contenu sémantique de stimuli avant de donner leur réponse sur l'effet ventriloque. Tant qu'il y avait un minimum de différence en azimut entre les stimuli sonore et visuel, le fait de poser des questions sémantiques n'avait aucun effet sur les réponses de fusion image/son des sujets. Dans le plan médian (où il n'y a aucune différence d'azimut), l'effet était significatif mais contraire à notre hypothèse de départ. En effet, nous avons pu observer que l'effet ventriloque fonctionnait moins bien lorsque le sujet devait se focaliser sur le contenu sémantique.

Les résultats suggèrent que :

— en azimut, l'efficacité de l'effet ventriloque dépend principalement de la position horizontale du stimulus sonore par rapport à la position du stimulus visuel. Comme les performances de localisation en azimut sont précises et comparables d'un individu à l'autre, l'efficacité de l'effet ventriloque est limitée et la variabilité inter-sujet est modérée par rapport à celle observée en élévation ;

— en élévation, la localisation auditive n'est pas précise, et l'influence de la position du stimulus sonore décroît substantiellement au profit d'autres facteurs à variabilité inter-individuelle élevée tels que l'attention du sujet et le réalisme de la combinaison son-image. Ainsi, des seuils plus importants sont obtenus (surtout si la combinaison son-image est convaincante) et la variabilité inter-sujet augmente.

Implications pour la cohérence audiovisuelle en élévation au cinéma

Les recommandations CST préconisent une distance minimale D_{min} entre le premier rang et un écran de largeur L telle que $D_{min} = 0,8 \times L$ (Recommandation RT 012, 2003).

Dans la Fig. 5.10, on considère une personne assise au premier rang avec ses oreilles situées à mi-hauteur de l'écran (on suppose que la diffusion se fait dans le plan horizontal, comme il est d'usage, avec les enceintes placées à mi-hauteur de l'écran). Il s'agit de la pire place possible, puisqu'elle maximise les disparités verticales susceptibles d'apparaître pendant le film. Si l'écran est au format 16/9, alors sa hauteur h est égale à $\frac{9}{16}L$. La disparité verticale maximale Ψ_{max} entre son et image est obtenue pour une source visuelle située soit au bord haut de l'écran, soit au bord bas (ce qui en soi constitue également un cas extrême et rare). Dans ce cas :

$$\Psi_{max} = \arctan \frac{\frac{h}{2}}{D_{min}} = \arctan \frac{\frac{9L}{32}}{0.8L} \approx 19.4°$$

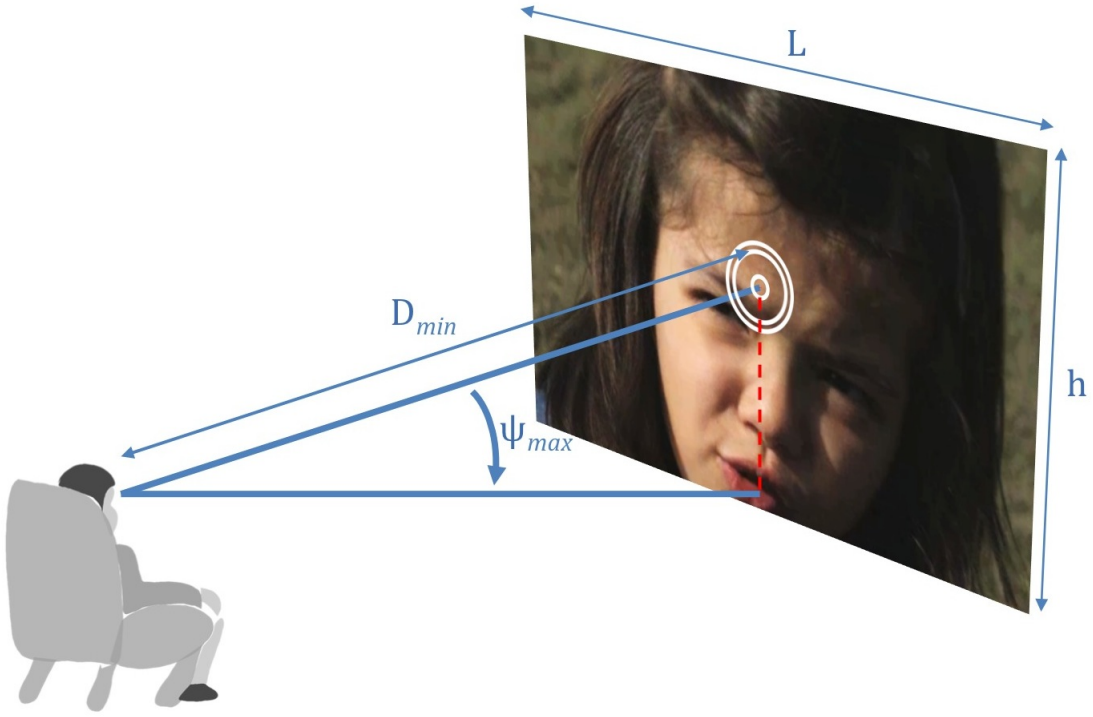


FIGURE 5.10 – Disparité verticale Ψ_{max} entre la bouche de l'actrice située au bord bas de l'écran et sa voix diffusée à mi-hauteur de l'écran.

Dans l'expérience IV, nous n'avons pas souhaité utilisé les « seuils à 50% » à 90° d'orientation δ car ces seuils ne pouvaient pas être estimés de manière fiable pour tous les sujets. Cette estimation est cependant fiable pour le sujet H, qui était le sujet le plus discriminant en élévation : à 90° d'orientation δ, son seuil à 50% était égal à 23°.

Cette valeur est supérieure à Ψ_{max}, ce qui montre que l'effet ventriloque en élévation continuera d'être efficace même dans le pire cas de figure : un spectateur très discriminant assis au premier rang et regardant une source visuelle située en bordure haute ou basse de l'écran. Ce phénomène a été démontré pour une source visuelle et une enceinte à azimut 0°, mais il est également valable à n'importe quel azimut. En effet, Werner *et al.* (2013) ont observé avec des sources visuelles et sonores à 30° d'azimut que les sujets étaient tout aussi tolérants vis-à-vis des disparités verticales, voire même encore plus tolérants, qu'à 0° d'azimut.

Cette constatation suggère que l'influence de la cohérence audiovisuelle en élévation sur la qualité d'expérience audiovisuelle est probablement négligeable dès lors qu'un mixage est déjà cohérent en azimut.

Le test n'ayant malheureusement pas pu être conduit avec l'image projetée en 2D (la durée du test, déjà importante, aurait été doublée), nos résultats ne peuvent s'appliquer qu'à une projection en 3D-s. Il est assez difficile à ce stade de prédire quels auraient été les résultats

avec une image en 2D :

— André *et al.* (2012) et Iljazovic *et al.* (2012) semblent suggérer que les sujets sont plus exigeants en termes de cohérence spatiale lorsque l'image est en 3D, car les attentes en termes de réalisme sonore sont plus importantes. Les seuils seraient donc plus élevés en 2D ;

— nous pourrions également supposer que puisque la stéréoscopie procure une plus grande sensation de présence (Ijsselsteijn *et al.*, 2001), les sujets auront plus la sensation d'être « en présence » de l'acteur et présumeront plus facilement que la bouche et la voix de l'acteur proviennent du même endroit. Les seuils seraient alors plus élevés en 3D.

5.2 Expérience V : Influence de la stéréoscopie sur l'appréciation de la cohérence audiovisuelle spatiale

5.2.1 Introduction

Nous avons vu dans le chapitre 1.2.3 que les ingénieurs du son en général ne spatialisaient pas les dialogues (et les effets sonores) et les envoyaient directement dans l'enceinte centrale (Toole, 2008), quelle que soit la position à l'écran de leur source visuelle correspondante. Une disparité peut donc apparaître entre une source sonore (par exemple, la voix de l'acteur) et sa source visuelle associée (la bouche de l'acteur) si la position à l'écran de la source visuelle ne coïncide pas avec l'enceinte centrale.

Certains justifient cette pratique par l'effet ventriloque : la cohérence audiovisuelle spatiale n'est pas indispensable puisque l'audience perçoit une source sonore au même endroit que sa source visuelle associée même quand les deux sources ne sont pas physiquement au même endroit. Cet argument est cependant discutable, puisque nous avons montré dans le chap. 3.1 et dans l'expérience IV que l'effet ventriloque ne fonctionnait plus lorsque l'écart entre la source sonore et la source visuelle associée était trop important. De plus, plusieurs ingénieurs du son et chercheurs ont suggéré que la cohérence audiovisuelle spatiale pouvait améliorer significativement l'expérience des spectateurs, surtout pour des contenus en 3D-s : il s'agit de notre sous-hypothèse 2, que nous souhaitons ici vérifier.

Michael Semanick, par exemple, affirme avoir plus latéralisé les dialogues pour la version 3D-s d'*Alice au pays des Merveilles* que pour la version 2D (Gambier, 2010). Paul Martin Smith, monteur du film d'Eric Brevig *Voyage au Centre de la Terre*, estime également qu'un mixeur devrait toujours travailler avec l'image projetée en 3D, car cela influence la spatialisation des sources sonores (Krohn, 2009). En revanche, les ingénieurs du son de *Hugo Cabret* ou d'*Avatar* considèrent que l'influence de la stéréoscopie est négligeable.

Nous avons également montré dans le chapitre 3.2 que des études précédentes avaient obtenu des résultats contradictoires concernant l'apport de la cohérence audiovisuelle spatiale à la qualité d'expérience audiovisuelle :

— dans l'étude d'André *et al.* (2012), la cohérence audiovisuelle en azimut et en profondeur avait un impact négatif sur la sensation de présence des sujets ;
— dans l'étude de Moulin (2015), la cohérence audiovisuelle en profondeur avait un impact positif sur la sensation d'immersion des sujets, mais pour un nombre très limité de séquences (2 sur 9) ;
— dans l'étude de Kruszielski *et al.* (2012) la cohérence audiovisuelle en profondeur améliorait l'adéquation du son à l'image pour toutes les séquences. Cependant, cette tendance n'était pas spécifique à la stéréoscopie, puisque l'amélioration était la même en 3D-s et en 2D.

Les séquences utilisées dans ces études étaient peu nombreuses ou peu variées en termes de contenu. Or, nous avons montré dans les expériences I et II que l'effet de la stéréoscopie pouvait fortement dépendre du contenu des séquences présentées (voir chap. 4). Nous proposons donc de réévaluer la pertinence de la cohérence audiovisuelle spatiale au cinéma avec 8 séquences 3D-s variées.

La cohérence audiovisuelle spatiale peut être étudiée selon 3 axes : en azimut, en élévation et en profondeur.

En azimut

Dans le chapitre 3.1 et dans l'expérience IV, nous avons montré que l'effet ventriloque ne fonctionnait plus si les disparités azimutales entre son et image devenaient trop importantes, même dans le cas d'une combinaison audiovisuelle « réaliste » : 20° avec Komiyama (1989), 18° avec André *et al.* (2014), 15° en moyenne lors de notre expérience IV, etc.

Or, les disparités azimutales au cinéma peuvent être plus importantes que ces seuils, surtout pour des sources visuelles projetées loin du centre de l'écran (et donc loin de leur correspondant sonore s'il est diffusé sur l'enceinte centrale). Nous pouvons donc supposer que l'effet ventriloque ne fonctionne pas en permanence au cinéma, et que les sujets risquent de détecter des disparités entre son et image qui pourraient altérer la qualité de leur expérience audiovisuelle.

Cependant, le genre de tâches que les études sur l'effet ventriloque impliquent focalisent l'attention du sujet sur la cohérence audiovisuelle spatiale, et met donc le sujet dans une situation bien éloignée de celle qu'il expérimente lorsqu'il va voir un film au cinéma :

— les sujets risquent de détecter des disparités spatiales dont ils ne se seraient pas aperçus si on ne leur avait pas posé la question (Komiyama, 1989) ;

— il est tout à fait possible que les sujets trouvent la bande-son adaptée à l'image quand bien même ils détectent des disparités. Par exemple, dans l'étude de Komiyama (1989), la moitié des sujets pouvaient détecter une disparité de 10°, mais cette disparité n'était jugée gênante qu'à partir de 30° ;

Il est donc difficile de prédire à partir des résultats d'études sur l'effet ventriloque quel effet aura telle ou telle incohérence audiovisuelle en azimut sur la qualité d'expérience audiovisuelle.

En élévation

Etudier en parallèle les effets simples de la cohérence audiovisuelle en azimut, de la cohérence audiovisuelle en élévation, de la cohérence audiovisuelle en profondeur, puis les différentes interactions entre les dimensions aurait rendu le test bien trop fastidieux et surtout trop long pour le sujet. Il fallait donc éliminer une des trois dimensions. Or, nous avons montré dans l'expérience IV que l'effet ventriloque fonctionnait toujours en élévation au cinéma,

même dans le pire cas de figure (un spectateur très discriminant assis au premier rang et regardant une source visuelle située en bordure haute ou basse de l'écran). Nous supposerons donc que l'influence de la cohérence audiovisuelle en élévation sur la qualité d'expérience audiovisuelle est négligeable (il faudrait cependant vérifier cette hypothèse : en effet, une source sonore disparate pourrait être localisée au même endroit que sa source visuelle correspondante tout en « sonnant » moins réaliste que lorsqu'elle est physiquement cohérente avec la source visuelle). Nous n'étudierons donc pas la cohérence audiovisuelle en élévation dans le cadre de la présente expérience.

En profondeur

Pour étudier l'influence de la cohérence audiovisuelle en profondeur, nous aurions pu envisager des systèmes de reproduction physique du champ sonore, tels que la Wave Field Synthesis (WFS) (Berkhout *et al.*, 1993) ou l'Ambisonics (Gerzon, 1973). Cependant, les indices binauraux n'apportent plus beaucoup d'informations pour des distances égocentriques supérieures à 1 m (Shinn-Cunningham, 2000) (surtout lorsque le sujet reste immobile, comme au cinéma), et les études de Moulin (2015) et d'André *et al.* (2012) suggèrent que l'impact des indices binauraux sur les impressions des sujets est limité (voir chap. 3.2).

Par contre, Kruszielski *et al.* (2012) ont obtenu un effet bien plus significatif de la cohérence en profondeur sur les impressions des sujets en se reposant plutôt sur les indices monauraux de la perception de la distance : intensité, rapport champ direct/champ diffus, coloration spectrale. Nous utiliserons donc des outils nous permettant de simuler ces indices monauraux (réglages de niveau, équalisation, réverbération artificielle et prises de son en champ diffus), ce qui nous permettra en plus de nous rapprocher des pratiques actuelles de mixage (voir chap. 1.2.3).

Présentation de l'expérience

Dans cette expérience, 8 séquences 3D-s ont été présentées au sujet. Pour chaque séquence, les sujets devaient évaluer à quel point la bande-son leur paraissait « adaptée » à l'image. Selon la bande-son, les sources sonores pouvaient être plus ou moins cohérentes en azimut et en profondeur avec la position de leur source visuelle respective sur l'écran. Les séquences étaient également présentées dans leur version 2D, afin de déterminer si la stéréoscopie avait un impact significatif sur les attentes des sujets en termes de spatialisation.

5.2.2 Matériel et méthode

Lieu de l'expérience

Le test s'est déroulé dans une salle du département « Image & Son » de l'Université de Brest spécialement conçue pour la post-production cinématographique. Les lumières étaient

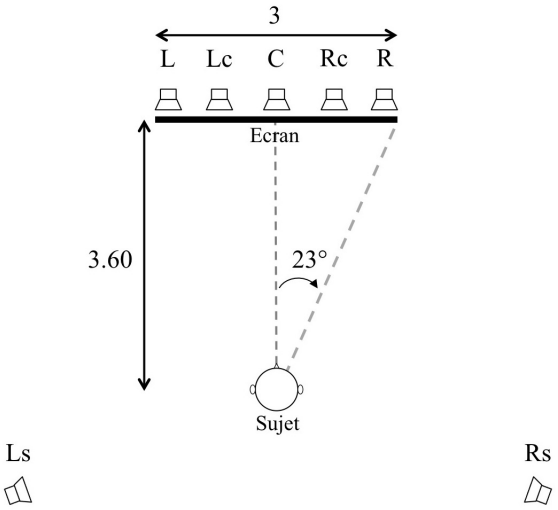


FIGURE 5.11 – Configuration de l'expérience V. Dimensions en mètres

éteintes et le sujet était assis à 3,60 m de l'écran (voir Fig. 5.11).

L'image (25 i/s) était projetée par un projecteur numérique 3D Epson EH-TW6000 (synchronisé avec des lunettes 3D actives Epson ELPGS01), sur un écran acoustiquement transparent (3 m de largeur). Le champ visuel des images projetées était de 46°, légèrement supérieur à l'angle idéal de 45° proposé par Dolby (1994).

La diffusion des stimuli et l'enregistrement des notes attribuées par les sujets étaient assurés par un logiciel programmé sous Max/MSP sur un ordinateur MacBook Pro relié à une interface RME MADIface USB.

Le système de diffusion était composé de 7 enceintes Amadeus PMX4, alimentées par un convertisseur numérique-analogique D.O.Tec Andiamo 2.DA et des amplificateurs Audac DPA154 :

— 5 enceintes étaient cachées derrière l'écran : Gauche (L), Gauche-droite (Lc), Centre (C), Droite-centre (Rc) et Droite (R) ;
— 2 enceintes étaient placées derrière le sujet pour la diffusion des « surround » : Surround gauche (Ls) et Surround droite (Rs).

Une telle disposition, avec 5 enceintes frontales, a été recommandée par de nombreux formats au cours des dernières décennies : Cinerama (1952), Cinemascope (1953), SDDS (1993), NHK 22.2 (toujours en développement), Dolby Atmos (2012), etc. (voir chap. 1.2).

Chaque enceinte était filtrée numériquement pour compenser les différences de timbre, et leur gain avait été calibré de manière à ce que la diffusion d'un bruit rose à - 20 dBfs RMS produise à la position d'écoute optimale un niveau de pression sonore égal à 85 dBC par enceinte. La fréquence d'échantillonnage des signaux audio était de 48 kHz et la quantification était effectuée sur 24 bits.

FIGURE 5.12 – Tournage de la séquence 2 de l'expérience V.

Séquence

8 séquences 3D-s ont été utilisées pour le test, ainsi que leur version 2D. 5 séquences extraites de productions professionnelles ont été utilisées, ainsi que trois séquences spécialement tournées pour le test avec une caméra 3D Panasonic AG-3DP1. Les séquences ont été choisies de manière à couvrir une large gamme de dynamiques (plans statiques, travelling avant, panoramique, etc.), de valeurs (plans serrés, plans moyens, plans larges, etc.), de décors (intérieur, plage, forêt, etc.) et de situations (dialogues, un homme en bicyclette, une femme faisant la vaisselle en écoutant la radio, etc.). Il est à noter que les expérimentateurs avaient assuré la prise de son, le montage et le mixage de la bande-son originale de 4 des 5 séquences extraites de productions professionnelles (voir Fig. 5.12).

Pour chaque séquence, le niveau avait été au préalable fixé subjectivement par les expérimentateurs, comme c'est souvent le cas dans les tests subjectifs (IEC 60268-13, 1998) et dans les cinémas (Recommandation RT 013, 2006).

Les séquences sont plus largement détaillées dans l'annexe H

Cohérence audiovisuelle en azimut

Chaque séquence était composée :

— d'objets sonores, tels que dialogues, sons de voiture, de bicyclette, craquements de branches, etc. Chaque objet était enregistré sur une piste monophonique individuelle qui pouvait être latéralisée n'importe où entre les enceintes ;

— d'ambiances 5.0 (ambiances diffuses de forêt, de mer, de vent, réverbération, etc.) diffusées dans les enceintes L, C, R, Ls et Rs (comme il est d'usage dans les productions professionnelles).

Pour chaque séquence, deux spatialisations différentes étaient proposées aux sujets :

— un mixage « classique », dans lequel les objets sonores étaient diffusés sur l'enceinte centrale ;
— un mixage « cohérent », c'est-à-dire avec les objets sonores reproduits au même azimut que la position à l'écran de leur correspondant visuel.

Dans les mixages « cohérents », les objets sonores étaient placés virtuellement en azimut à l'aide d'un panning d'amplitude sur les 5 enceintes placées derrière l'écran utilisant une loi de panning dite "tangente" (Pulkki et Karjalainen, 2001). Il n'y avait pas d'objet hors-champ dans les séquences : les enceintes « surround » n'étaient donc pas concernées et servaient uniquement pour la diffusion des ambiances diffuses. Le choix s'est porté sur un panning d'amplitude car la plupart des cinémas dans le monde sont équipés avec des systèmes se basant sur des lois d'intensité pour la spatialisation des sources.

À l'aide d'un logiciel programmé sous Max/MSP, la position des objets visuels (les bouches des personnages, et autres objets tels que voiture, bicyclette, etc.) a été « trackée » image par image pour chaque séquence. A partir de la largeur de l'écran (3 m) et de la distance entre le sujet et l'écran (3,60 m), les positions mesurées ont été converties en positions angulaires et envoyées dans un logiciel de spatialisation également développé sous Max/MSP. Des fichiers audio de 5 pistes ont ensuite été générés pour alimenter les enceintes derrière l'écran avec les gains adéquats permettant aux objets sonores d'être localisés aux azimuts désirés. Les ambiances 5.0 n'étaient pas affectées par ces opérations de spatialisation et étaient diffusées sur les enceintes L, C, R, Ls et Rs comme pour les mixages « classiques ».

Simulation de la profondeur

Pour chaque séquence, les expérimentateurs ont produit deux versions différentes concernant la profondeur sonore :

— un mixage « proximité », sans simulation de la profondeur : les prises de son de proximité originales des objets ont été utilisées telles quelles. Aucun effort d'aucune sorte n'a été fait pour produire un son cohérent avec la profondeur des objets visuels ;
— un mixage « distance simulée » : Pour 5 séquences (séquences 1 à 5), des traitements numériques utilisés en pratique par les professionnels (équalisation et réverbération) ont été appliqués sur les objets afin d'atténuer la sonorité « prise de son de proximité » des enregistrements originaux et d'être plus respectueux de la position en profondeur des objets visuels correspondants. Le niveau sonore a également été baissé pour les objets visuels lointains. Pour les 3 autres séquences (séquences 6 à 8), des prises de son enregistrées au tournage en champ diffus ont été utilisées. Comme pour la cohérence en azimut, ces simulations de profondeur ne concernaient ni les ambiances 5.0 ni les enceintes surround.

Séquence	ΔL (dB RMS)
1	5.0
2	6.9
3	7.2
4	1.8
5	0.89
6	4.9
7	1.3
8	1.2

TABLEAU 5.2 – Différences de niveau sonore moyen des objets dans le mixage « proximité » et dans le mixage « distance simulée » (dB RMS) pour chaque séquence. Une valeur positive signifie que les objets du mixage « proximité » sont plus forts que les objets du mixage « distance simulée » associé.

Version	Mode Visuel	Cohérence Azimutale	Simulation de la profondeur
1	2D	Non (mixage « classique »)	Non (mixage « proximité »)
2	2D	Non (mixage « classique »)	Oui (mixage « distance simulée »)
3	2D	Oui (mixage « cohérent »)	Non (mixage « proximité »)
4	2D	Oui (mixage « cohérent »)	Oui (mixage « distance simulée »)
5	3D-s	Non (mixage « classique »)	Non (mixage « proximité »)
6	3D-s	Non (mixage « classique »)	Oui (mixage « distance simulée »)
7	3D-s	Oui (mixage « cohérent »)	Non (mixage « proximité »)
8	3D-s	Oui (mixage « cohérent »)	Oui (mixage « distance simulée »)

TABLEAU 5.3 – Les 8 différentes versions pour chaque séquence. Les parenthèses indiquent comment il sera fait référence à chacune des conditions dans la suite du texte.

Le tableau 5.2 présente pour chaque séquence les différences de niveau sonore moyen des objets (dialogues et effets *in*) résultant de ces simulations de profondeur. Pour certaines séquences (5, 7 et 8), la différence de niveau est faible, car les différences entre mixages « proximité » et mixages « distance simulée » concernent plutôt la réverbération et la balance spectrale. Par exemple, la Fig. 5.13 permet de comparer les balances spectrales des objets du mixage « proximité » avec ceux des mixages « distance simulée » dans la séquence 8.

Un test préliminaire informel sans image (passé par les expérimentateurs ainsi que 4 sujets naïfs) a confirmé que les différences entre mixages « proximité » et mixages « distance simulée » donnaient lieu pour chaque séquence à des différences de perception importantes.

Il y avait donc 8 versions différentes pour chacune des 8 séquences (résumées dans le Tableau 5.3), ce qui veut dire que les sujets devaient évaluer en tout $8 \times 8 = 64$ stimuli différents.

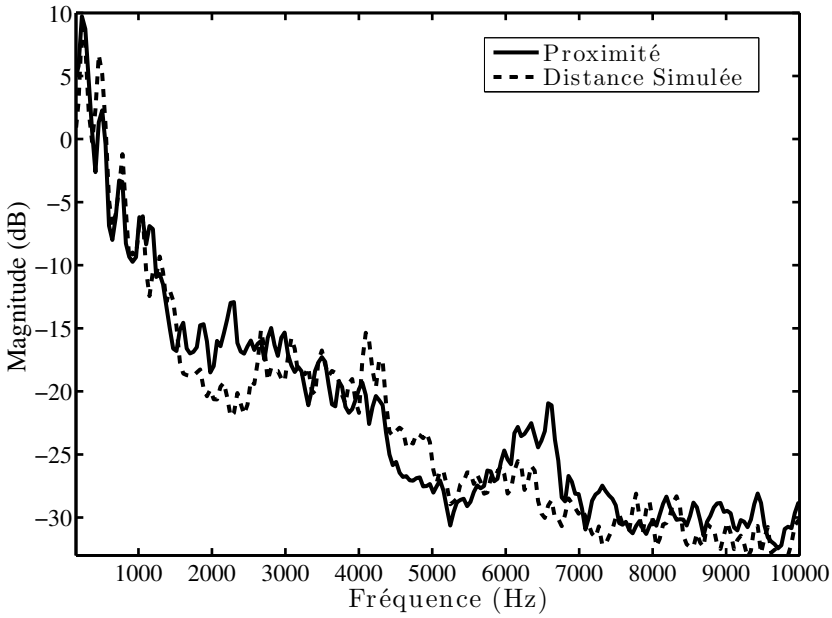


FIGURE 5.13 – Balance spectrale des objets du mixage « proximité » (ligne continue) comparée avec la balance spectrale des objets du mixage « distance simulée » (ligne en pointillés) dans la séquence 8. Pour cette séquence, le mixage « proximité » a été obtenu à partir de prises de son de proximité, tandis que le mixage « distance simulée » a été obtenu à partir de prises de son enregistrées au tournage en champ diffus. Cette figure montre que les mixages « proximité » et mixages « distance simulée » de certaines séquences diffèrent plus par leur balance spectrale que par leur niveau global.

Protocole

16 sujets naïfs ont pris part à l'expérience (7 hommes, 9 femmes, âgés de 16 à 50 ans). Tous les sujets allaient régulièrement au cinéma (au moins une fois par mois) et avaient déjà vu au moins un film en 3D au cinéma. Par contre, ils n'avaient jamais participé à des tests perceptifs. Ils étaient rémunérés pour leur participation et ont tous déclaré avoir une vision et une audition normales.

Après chaque présentation d'un stimulus, les sujets devaient répondre à la question « À quel point jugez-vous le son de cette séquence adapté à l'image ? », en déplaçant un curseur le long d'une échelle à 100 points affichée sur un écran d'ordinateur, dont les extrémités étaient labellisées « pas adapté du tout » (correspondant à la note 0/100) et « très adapté » (correspondant à la note 100/100). Une image de l'interface graphique est montrée dans la Fig. 5.14.

Cette échelle est similaire à celles utilisées par Kruszielski *et al.* (2012) et Kamekawa *et al.* (2011) (il est à noter qu'aucune recommandation n'existe pour l'instant pour de telles évaluations de contenus audiovisuels 3D-s (Moulin *et al.*, 2013)). Il n'y avait aucune gradation sur l'échelle ni aucun label intermédiaire prédéfini, afin d'éviter l'introduction de biais indésirables (Poulton, 1992; Zielinski *et al.*, 2008). Lorsque le sujet était satisfait de sa réponse, il devait cliquer sur un bouton « suivant » et un nouveau stimulus était alors présenté.

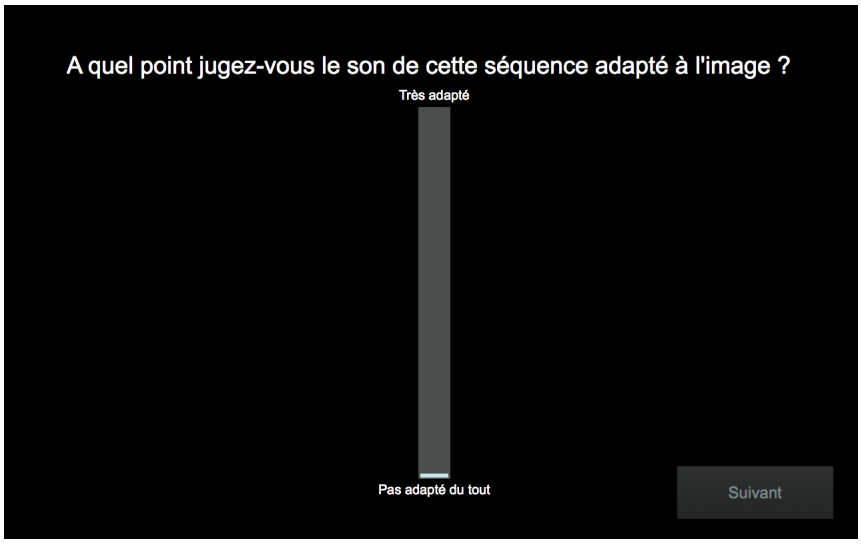


FIGURE 5.14 – Interface graphique de réponse pour l'expérience V.

Nous avons décidé de poser une question appelant à un jugement global et de ne pas focaliser l'attention du sujet sur la dimension spatiale. En effet, il a été montré dans le chap. 3.1.3 que focaliser l'attention sur la dimension spatiale pouvait rendre le sujet exagérément discriminant par rapport à la vraie vie. Par exemple, lorsque Komiyama (1989) a demandé à ses sujets de quantifier à quel point les disparités entre son et image des séquences présentées étaient perceptibles et gênantes, plusieurs sujets ont rapporté qu'ils n'auraient probablement pas remarqué ces disparités si on ne leur avait pas posé la question. Un jugement global nous paraissait donc plus écologiquement valide. Si les sujets ne détectent pas que certaines séquences sont spatialement cohérentes et d'autres non, et si la cohérence audiovisuelle spatiale n'influence par leur jugement global, cela tendra à montrer que l'apport de la cohérence audiovisuelle au cinéma est bien trop subtil pour être véritablement pertinent.

Les 64 stimuli à évaluer étaient présentés aléatoirement à tous les sujets et dans un ordre différent pour chaque sujet. Les sujets étaient priés de garder leurs lunettes 3D pendant la totalité de l'expérience. Chaque sujet a passé le test 2 fois, avec une pause de 15 minutes entre les deux sessions. La première session était précédée d'un pré-test de 5 minutes pour familiariser le sujet avec la tâche, le contenu des 8 séquences ainsi que l'interface graphique de réponse. La durée moyenne d'une session était d'environ 30 minutes (chaque séquence durait 20 secondes). Aucun sujet ne s'est plaint d'avoir ressenti de la fatigue visuelle ou de l'inconfort pendant le test.

Dans la suite du texte, nous emploierons l'expression *adéquation à l'image* pour parler d'une bande-son plus ou moins adaptée à l'image.

5.2.3 Résultats

Les données brutes ont subi une transformée en z, afin de minimiser les différences inter-sujets d'utilisation de l'échelle (ITU-R BS.1116-1, 1994; ITU-R BS.1286, 1997).

Vérification des hypothèses pour une ANOVA

Pour pouvoir utiliser une ANOVA à mesures répétées légitimement, deux hypothèses doivent être validées : l'hypothèse de normalité et celle de sphéricité (Howell, 2009).

Un test de Kolmogorov-Smirnov a été effectué sur chacune des cellules (niveau de signification à 5%) pour vérifier l'hypothèse de normalité. Le test a rejeté l'hypothèse d'une distribution normale pour 25 cellules sur 128. Cependant, de nombreuses études rapportent que l'ANOVA peut être robuste face à ce genre de violations (Hays, 1994), surtout si la taille d'échantillon est supérieure à 15 observations par cellule (Green et Salkind, 2013). Avec 16 observations par cellule, l'utilisation d'une ANOVA reste donc envisageable dans le cadre de notre étude.

Un test de Mauchly a montré que l'hypothèse de sphéricité était violée pour l'interaction « Répétition * Séquence * Mode Visuel * Cohérence Azimutale ». La valeur F de ce facteur a donc dû être corrigée en utilisant la correction de Greenhouse-Geisser (Girden, 1991).

Nous avons décidé de ne pas exclure le sujet n°8 de nos données car :

— Une discussion avec le sujet n'a pas permis de déceler de raisons valables pouvant justifier son exclusion (pas de problème de vue ou d'audition, pas d'intolérance à la 3D-s, pas de mauvaise compréhension apparente de la question posée etc.) ;
— Exclure le sujet ne modifiait pas les significativités des différents facteurs de l'ANOVA.

ANOVA

Les résultats sont présentés dans le Tableau 5.4, avec les effets et interactions des 5 facteurs suivants :

— S : Séquence (8 niveaux) ;
— V : Mode Visuel (2 niveaux : 2D vs 3D-s) ;
— A : Cohérence Azimutale (2 niveaux : « classique » vs. « cohérent ») ;
— D : Simulation de la Profondeur (2 niveaux : « proximité » vs. « distance simulée ») ;
— R : Répétition (2 niveaux) ;

Effet du Mode Visuel : 3D-s vs. 2D

Les résultats montrent que l'effet global du Mode Visuel n'a pas été significatif ($V : F(1, 15) = 0.314$, $p = 0.584 > 0.05$). Aucune interaction impliquant le Mode Visuel n'a été non plus significative. La stéréoscopie n'a donc influencé d'aucune façon les jugements des sujets sur l'adéquation du son à l'image.

Source	SS	DS	MF	F	Sig. p
R	162.348	1	162.348	0.055	0.817
S	111256.685	7	15893.812	8.055	0
V	252.329	1	252.329	0.314	0.584
A	61867.325	1	61867.325	8.938	0.009
D	3845.412	1	3845.412	4.661	0.047
R * S	1317.456	7	188.208	0.405	0.897
R * V	14.353	1	14.353	0.029	0.867
S * V	3253.707	7	464.815	0.855	0.545
R * A	7108.683	1	7108.683	6.379	0.023
S * A	15627.422	7	2232.489	3.327	0.003
V * A	99.911	1	99.911	0.295	0.595
R * D	732.89	1	732.89	0.983	0.337
S * D	69029.918	7	9861.417	8.692	0
V * D	728.12	1	728.12	1.473	0.244
A * D	1328.132	1	1328.132	3.903	0.067
R*S*V	4902.466	7	700.352	1.565	0.154
R*S*A	4870.679	7	695.811	1.656	0.128
R*V*A	21.475	1	21.475	0.044	0.837
S*V*A	2064.94	7	294.991	0.778	0.608
R*S*D	2836.195	7	405.171	1.143	0.342
R*V*D	3.457	1	3.457	0.009	0.925
R*A*D	220.755	1	220.755	0.352	0.562
S*A*D	1590.283	7	227.183	0.551	0.793
V*A*D	393.105	1	393.105	1.186	0.293
S*V*D	1478.64	7	211.234	0.353	0.927
R*S*V*A	3705.17	4.271	867.53	1.102	0.365
R*S*V* D	1270.185	7	181.455	0.326	0.941
R*S*A*D	4048.054	7	578.293	1.325	0.246
R*V*A*D	538.303	1	538.303	1.873	0.191
S*V*A*D	2700.787	7	385.827	0.768	0.616
R*S*V*A*D	1435.954	7	205.136	0.422	0.887

Tableau 5.4 – Résultats de l'ANOVA pour l'expérience V.

Effet de la Cohérence Azimutale : « classique » vs. « cohérent »

Les résultats montrent qu'il y a eu un effet global significatif de la Cohérence Azimutale (A : $F(1, 15) = 8.938$, $p = 0.009 < 0.01$), avec des moyennes respectivement égales à 71 et 60 pour les mixages « cohérents » et « classiques ». Les sujets ont donc trouvé que la Cohérence Azimutale pouvait améliorer significativement l'adéquation du son à l'image.

L'interaction Séquence * Cohérence Azimutale a également été significative (S*A : $F(7, 105) = 3.327$, $p = 0.003 < 0.01$), ce qui signifie que l'effet de la Cohérence Azimutale sur les jugements des sujets était différent selon la séquence. Un post-hoc LSD de Tukey, comparant l'effet de la Cohérence Azimutale séquence par séquence, a révélé que la Cohérence améliorait l'adéquation de la bande-son pour 5 séquences : séquence 1 ($p = 0.003$), séquence 2 ($p = 0.010$), séquence 3 ($p = 0.003$), séquence 6 ($p = 0.049$) et séquence 7 ($p = 0.004$). Par contre, la Cohérence n'a pas eu d'effet pour la séquence 4 ($p = 0.097$), la séquence 5 ($p = 0.884$) ainsi

que la séquence 8 (p = 0.176).

Les notes moyennes avec leur intervalle de confiance à 95%, obtenues pour la Cohérence Azimutale séquence par séquence, sont présentées dans la Fig. 5.15.

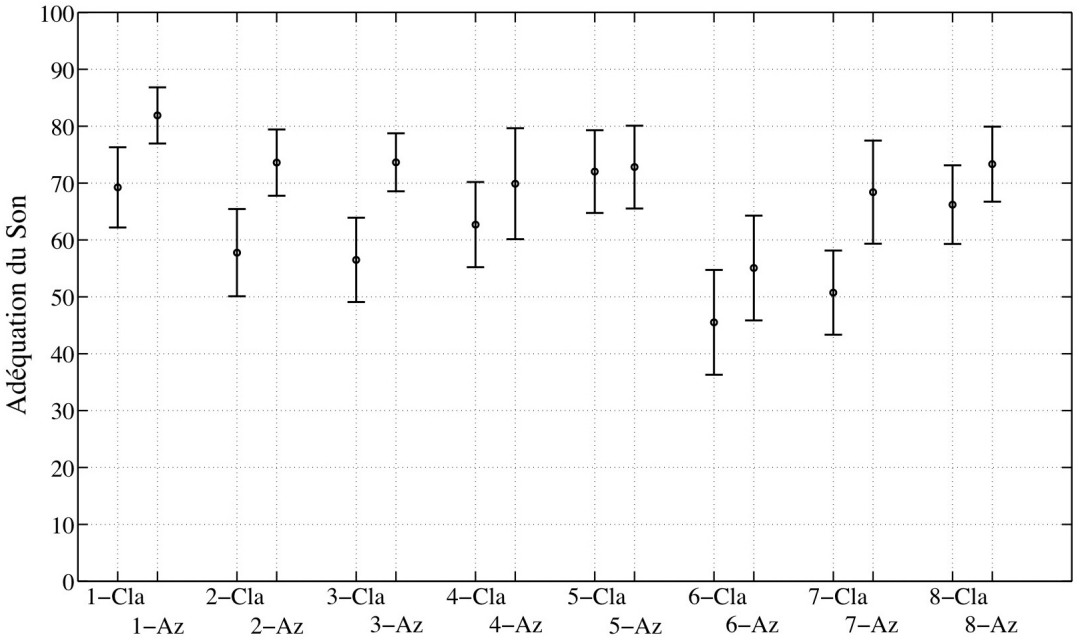


FIGURE 5.15 – Effet de la Cohérence Azimutale (« classique » vs. « cohérent ») sur l'adéquation du son à l'image pour chaque séquence.
Cla. : mixage « classique » avec objets diffusés sur l'enceinte centrale.
Az. : mixage « cohérent » en azimut.
Notes moyennes avec intervalles de confiance à 95%.

Effet de la Simulation de la Profondeur : « proximité » vs. « distance simulée »

Les résultats montrent qu'il y a eu un effet global significatif de la Simulation de la Profondeur (D : F(1, 15) = 4.661, p = 0.047 < 0.05), avec des moyennes égales à 67 pour les mixages « distance simulée » (avec traitements numériques ou prises de son en champ diffus) et 64 pour les mixages « proximité » (utilisation « brute » des prises de son en proximité originales). Les sujets ont donc trouvé que la simulation de la profondeur pouvait significativement améliorer l'adéquation du son. Cependant, cet effet n'est pas aussi prononcé que celui observé avec la Cohérence Azimutale.

L'interaction « Séquence * Simulation de la profondeur » a également été significative (S*D : F(7, 105) = 8.692, p < 0.001), ce qui signifie que l'effet de la Simulation de la Profondeur était différent selon la séquence. Un post-hoc LSD de Tukey, comparant l'effet de la Simulation de la Profondeur séquence par séquence, a montré que, malgré des différences importantes entre mixages « proximité » et mixages « distance simulée » (par exemple, $\Delta L = 7.9$ dB pour la séquence 3, voir Tableau 5.2), la Simulation de la Profondeur n'avait eu d'effet sur l'adéquation du son que pour une seule séquence : la séquence 6 (un homme poussant un chariot dans un couloir, $p < 0.001$).

Les notes moyennes avec leur intervalle de confiance à 95%, obtenues pour la Simulation de la Profondeur séquence par séquence, sont présentées dans la Fig. 5.16.

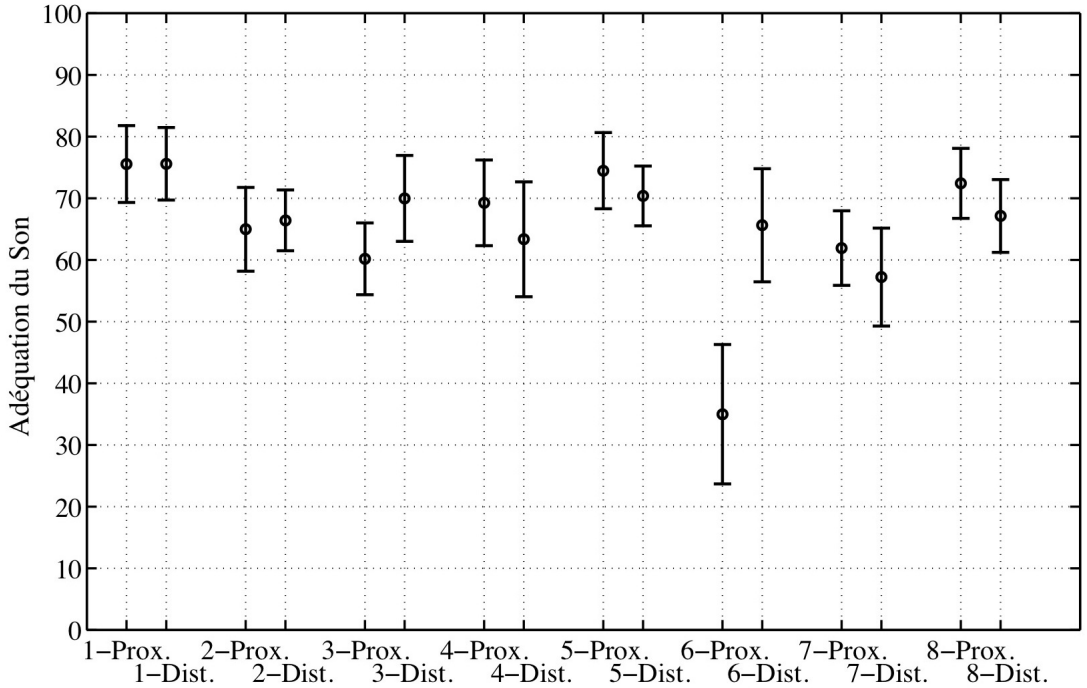


FIGURE 5.16 – Effet de la Simulation de la Profondeur (« proximité » vs. « distance simulée ») sur l'adéquation du son à l'image pour chaque séquence.
Prox. : mixage « proximité », sans simulation de la profondeur.
Dist. : mixage « distance simulée ».
Notes moyennes avec intervalles de confiance à 95%.

Effet de la Répétition

L'effet principal de la Répétition et les interactions impliquant la Répétition n'ont pas été significatifs, sauf l'interaction « Répétition * Cohérence Azimutale » (R*A : $F(1, 15) = 6.379$, $p = 0.023 < 0.05$). Un post-hoc LSD de Tukey a montré que la Cohérence Azimutale avait amélioré l'adéquation du son lors de la seconde session du test ($p = 0.002$), mais n'avait eu aucun impact sur les jugements des sujets pendant la première session ($p = 0.085 > 0.05$).

Les notes moyennes avec leur intervalle de confiance à 95%, obtenues pour la Cohérence Azimutale dans les deux sessions, sont présentées dans la Fig. 5.17.

La Fig. 5.18 montre les notes moyennes avec leur intervalle de confiance à 95%, obtenues pour la Cohérence Azimutale séquence par séquence dans la session 2. L'amélioration apportée par la cohérence azimutale est plus marquée et semble concerner un plus grand nombre de séquences que lorsque l'analyse porte sur l'ensemble des sessions du test.

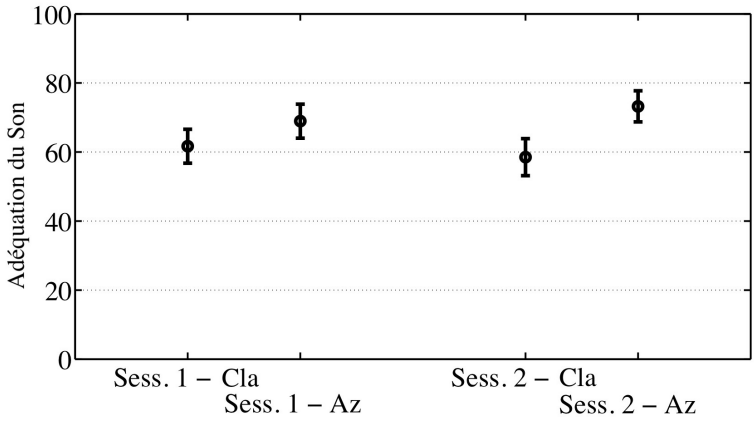


FIGURE 5.17 – Effet de la Cohérence azimutale (« classique » vs. « cohérent ») sur l'adéquation du son à l'image pour chaque session.
Cla. : mixage « classique » avec objets diffusés sur l'enceinte centrale.
Az. : mixage « cohérent » en azimut.
Notes moyennes avec intervalles de confiance à 95%.

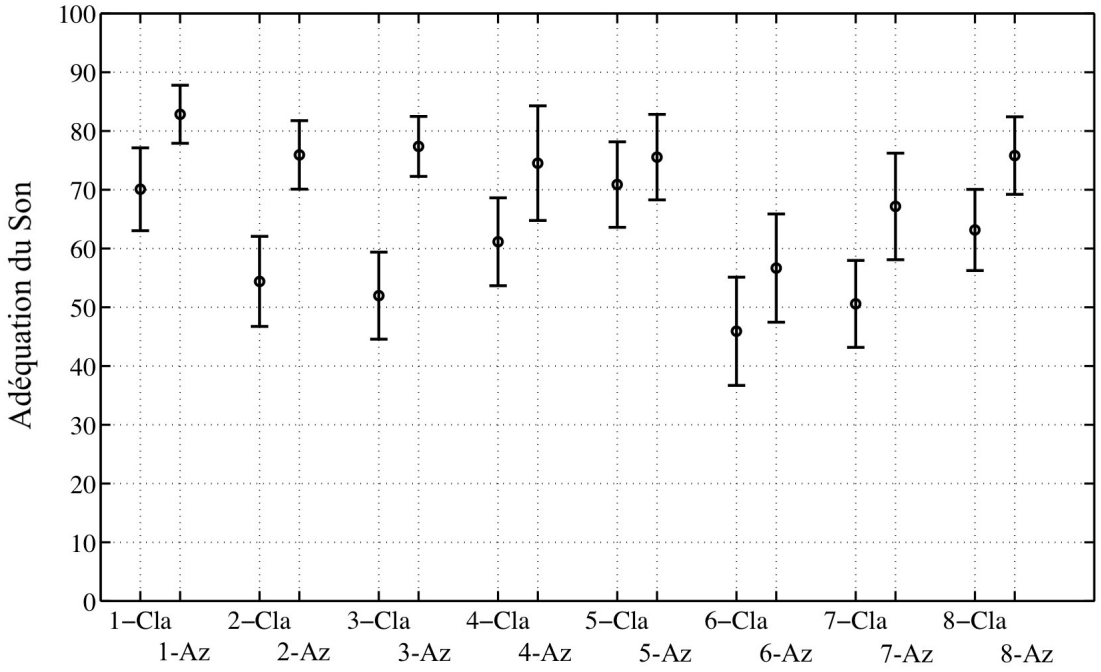


FIGURE 5.18 – Effet de la Cohérence Azimutale (« classique » vs. « cohérent ») sur l'adéquation du son à l'image pour chaque séquence de la session 2.
Cla. : mixage « classique » avec objets diffusés sur l'enceinte centrale.
Az. : mixage « cohérent » en azimut.
Notes moyennes avec intervalles de confiance à 95%.

5.2.4 Recherche de corrélations

Corrélations entre les caractéristiques spatiales des objets et l'amélioration apportée par la cohérence azimutale sur l'ensemble des sessions

Notre hypothèse pour les corrélations est simple et intuitive : l'amélioration apportée par la cohérence azimutale est d'autant plus élevée qu'elle concerne des objets ayant un azimut

élevé (et qui sont donc loin de l'enceinte centrale).

Nous avons décidé dans un premier temps de ne retenir pour chaque séquence que les objets étant le plus susceptibles de générer des différences de perception entre mixages « classiques » et « cohérents ». Dans la séquence 3 par exemple, il ne semble pas judicieux de prendre en compte le cycliste, dont les azimuts moyen et maximal égalent respectivement 0.6° et 1.4° : étant quasiment coïncident avec l'enceinte centrale, cet objet n'influencera certainement pas les différences de préférence entre mixages « classique » et « cohérent ».

Le Tableau 5.5 montre pour chaque séquence l'objet retenu avec ses caractéristiques spatiales. Ces caractéristiques sont ensuite comparées avec l'amélioration moyenne apportée par la cohérence en azimut par rapport à un mixage « classique » sur enceinte centrale (dernière colonne) pour les deux sessions (nous nous focaliserons ensuite uniquement sur les résultats de la seconde session). Nous rappelons que les caractéristiques spatiales des objets données sont celles des objets sonores lors des mixages « cohérents ». Bien évidemment, azimuts et vitesses des objets sonores sont toujours nuls dans le cas des mixages « classiques » sur enceinte centrale.

Séq.	Objet retenu	Azimut moyen	Azimut max	Vitesse moyenne	Vitesse max	Amélioration mixage cohérent
1	Homme	12.5°	23°	3.7°/s	16.8°/s	+12.6
2	Voiture	11.6°	23°	4.9°/s	22.7°/s	+15.8
3	Barque	16.4°	23°	4.4°/s	10°/s	+17.2
4	Enfant	12.1°	15.2°	1.1°/s	3.7°/s	+7.2
5	Branches	20.1°	23°	1.4°/s	29.3°/s	+0.8
6	Chariot	14.9°	23°	1.1°/s	4.8°/s	+9.5
7	Radio	18°	18°	0°/s	0°/s	+17.7
8	Marteau	11.0°	11.0°	0°/s	0°/s	+7.1

Tableau 5.5 – Caractéristiques spatiales des objets sonores des mixages « cohérents » et Amélioration apportée par la Cohérence Azimutale pour chaque séquence.

La Fig. 5.19 montre l'amélioration moyenne apportée par la cohérence en azimut en fonction de l'azimut moyen des objets retenus pour chaque séquence.

L'azimut semble plutôt bien expliquer les améliorations observées : l'amélioration augmente avec l'azimut moyen, sauf pour la séquence 5 (entourée en rouge), pour laquelle l'amélioration est la plus faible malgré un azimut moyen important. L'objet qui avait été retenu pour la séquence 5 était l'objet « Branches » (frappées l'une contre l'autre, puis jetées sur l'homme, avec des bruits d'impact sur la tête de l'homme puis sur le sol). Cependant, les impacts de branches étaient extrêmement brefs et ponctuels, ce qui explique probablement pourquoi ils n'ont pas influencé le jugement global des sujets. En retenant plutôt l'objet « Homme » pour cette séquence, on voit bien sur la Fig. 5.20 que l'azimut moyen semble bien mieux prédire l'amélioration observée. D'ailleurs, le coefficient de corrélation ρ de Spearman confirme que la corrélation est bonne ($\rho = 0.833$, p = 0.005).

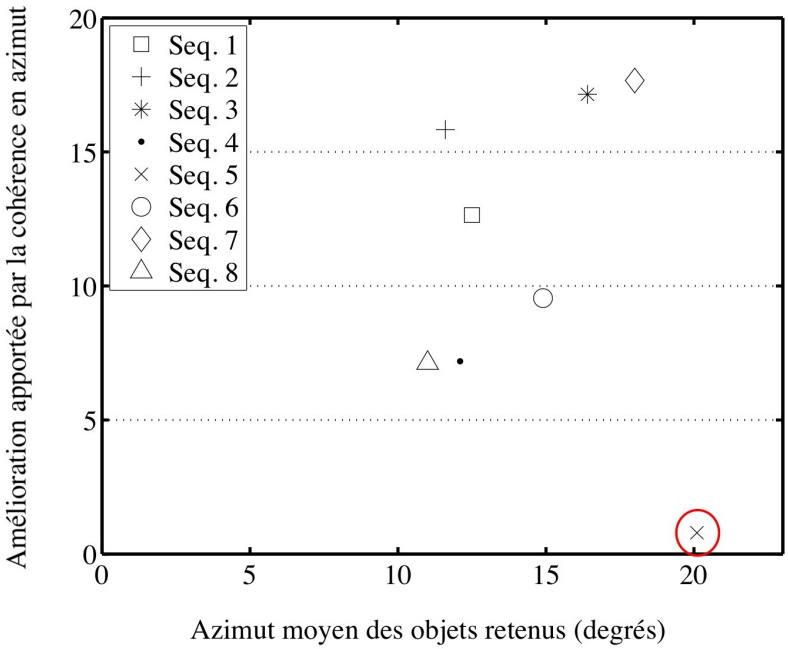


FIGURE 5.19 – Amélioration moyenne apportée par la cohérence en azimut en fonction de l'azimut moyen des objets retenus pour chaque séquence.

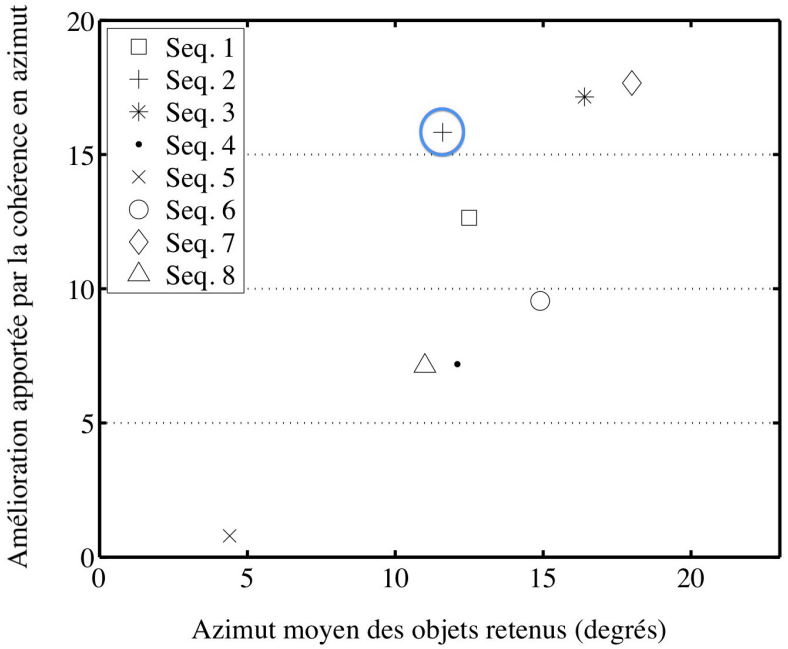


FIGURE 5.20 – Amélioration moyenne apportée par la cohérence en azimut en fonction de l'azimut moyen des objets retenus pour chaque séquence. Dans la séquence 5, l'objet « Homme » a été retenu plutôt que l'objet « branche », jugé trop bref et trop ponctuel.

L'amélioration pour la séquence 2 (entourée en bleu) est plus importante que ce que l'azimut de l'objet « Voiture » prédirait : il est probable que les sujets aient été fortement impressionnées par le virage violent qu'effectue la voiture à un moment (provoquant dans le mixage « cohérent » un déplacement rapide gauche-droite de l'objet sonore), comme le suggère les témoignages des sujets à l'issue du test. Il faudra donc également explorer les corrélations

avec les vitesses des objets.

La Fig. 5.21 montre l'amélioration moyenne apportée par la cohérence en azimut en fonction de l'azimut maximal des objets retenus pour chaque séquence.

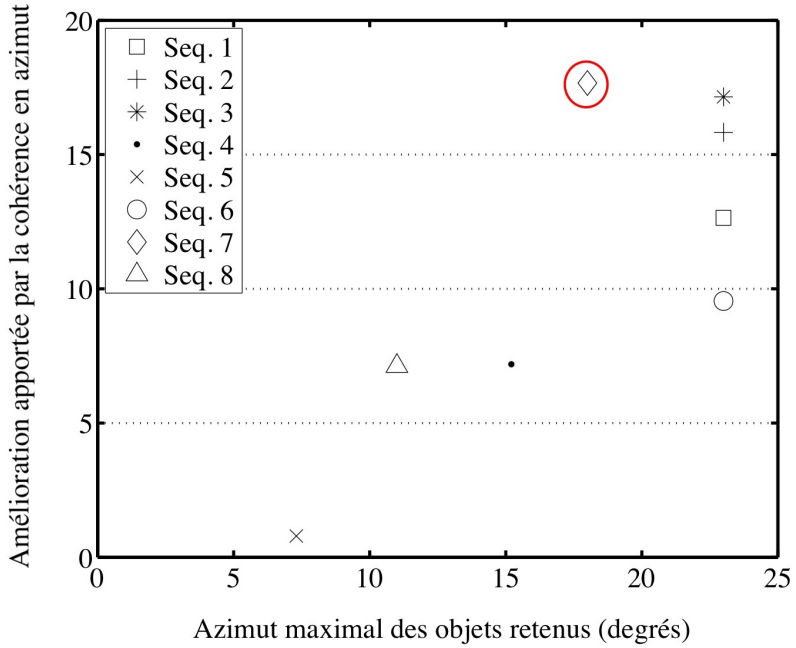


FIGURE 5.21 – Amélioration moyenne apportée par la cohérence en azimut en fonction de l'azimut maximal des objets retenus pour chaque séquence.

L'azimut maximal semble être un moins bon prédicateur pour l'amélioration ($\rho = 0.685$, p = 0.030). La séquence 7 (entourée en rouge), par exemple, a un azimut maximal égal à 18° et a pourtant une amélioration supérieure à celle observée dans les séquences 1, 2, 3 et 6, dont l'azimut maximal atteint 23°. Ce phénomène s'explique simplement : l'azimut de l'objet de la séquence 7 est en permanence égal à 18° alors que les objets des séquences 1, 2, 3 et 6 n'atteignent la valeur 23° que très ponctuellement et brièvement. L'azimut moyen semble donc être un meilleur prédicateur car il intègre la globalité de la séquence.

La Fig. 5.22 montre l'amélioration moyenne apportée par la cohérence en azimut en fonction de la vitesse moyenne des objets retenus pour chaque séquence.

La vitesse moyenne semble être un mauvais prédicateur de l'amélioration apportée par la cohérence en azimut ($\rho = 0.229$, p = 0.293). La séquence 7 (entourée en rouge), par exemple, a une vitesse nulle mais c'est pourtant la séquence pour laquelle la plus grande amélioration est observée. On peut cependant remarquer que :

— Si la vitesse moyenne est élevée (supérieure ou égale à 3.7°/s), alors l'amélioration sera significative (séquences 1, 2 et 3) ;

— Si la vitesse moyenne est faible (inférieure ou égale à 1.1°/s), alors on ne peut rien dire à première vue ;

La corrélation est encore plus mauvaise avec la vitesse maximale ($\rho = 0.192$, p = 0.325).

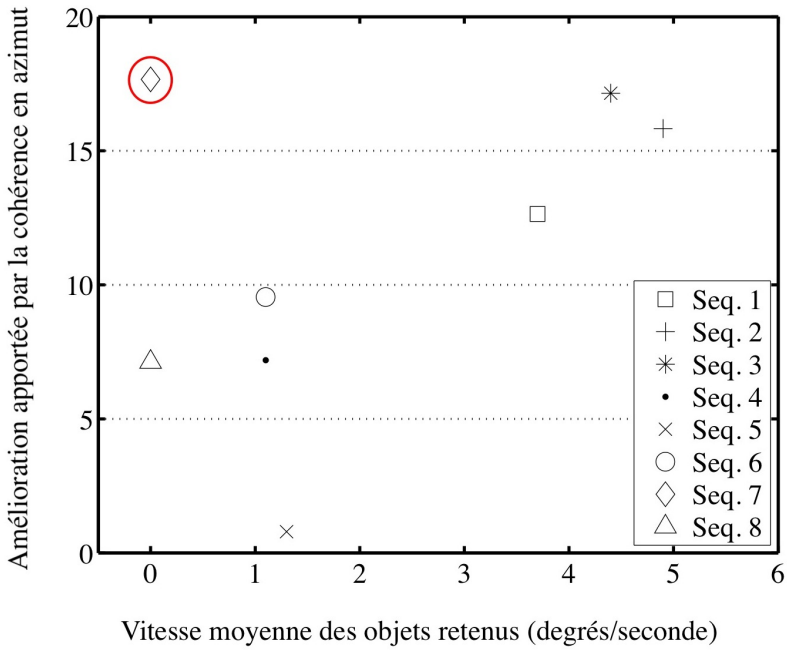


FIGURE 5.22 – Amélioration moyenne apportée par la cohérence en azimut en fonction de la vitesse moyenne des objets retenus pour chaque séquence.

Corrélations entre les caractéristiques spatiales des objets et la préférence pour la cohérence azimutale uniquement sur la seconde session

Les sujets ont été plus discriminants dans la seconde session que dans la première session pour la cohérence azimutale. Ce phénomène pourrait traduire un phénomène d'apprentissage et d'habituation à la cohérence azimutale. En effet, les sujets étaient peu habitués aux tests perceptifs et à l'écoute de sources spatialisées : peut-être ont-ils eu besoin d'un certain temps pour s'habituer et pour « calibrer » leurs attentes vis-à-vis de l'échelle de notation. Nous recherchons donc ici des corrélations uniquement avec les résultats obtenus dans la deuxième session du test.

Le Tableau 5.6 montre pour chaque séquence l'objet retenu avec ses caractéristiques spatiales. Ces caractéristiques sont ensuite comparées avec l'amélioration moyenne apportée par la cohérence en azimut par rapport à un mixage « classique » lors de la seconde session.

La Fig. 5.23 montre l'amélioration moyenne apportée par la cohérence en azimut en fonction de l'azimut moyen des objets lors de la seconde session.

L'azimut moyen semble assez bien prédire l'amélioration pour les séquences 1, 4, 5, 6, 7 et 8. Par contre, l'amélioration observée lors du test pour les séquences 2 et 3 est plus importante que ce que leur azimut moyen respectif laisserait présager. Ces résultats s'expliquent probablement par le fait que les objets des séquences 2 et 3 ont des vitesses moyennes élevées.

La Fig. 5.24 montre l'amélioration moyenne apportée par la cohérence en azimut en fonction de la vitesse moyenne des objets lors de la seconde session.

La vitesse moyenne ne semble pas à première vue un bon prédicateur de l'amélioration

Séq.	Objet retenu	Azimut moyen	Azimut max	Vitesse moyenne	Vitesse max	Amélioration mixage cohérent
1	Homme	12.5°	23°	3.7°/s	16.8°/s	+12.8
2	Voiture	11.6°	23°	4.9°/s	22.7°/s	+21.5
3	Barque	16.4°	23°	4.4°/s	10°/s	+25.4
4	Enfant	12.1°	15.2°	1.1°/s	3.7°/s	+13.4
5	Homme	4.4°	7.3°	1.3°/s	6.3°/s	+4.7
6	Chariot	14.9°	23°	1.1°/s	4.8°/s	+10.8
7	Radio	18°	18°	0°/s	0°/s	+16.6
8	Marteau	11.0°	11.0°	0°/s	0°/s	+12.7

TABLEAU 5.6 – Caractéristiques spatiales des objets sonores des mixages « cohérents » et amélioration apportée par la Cohérence Azimutale pour chaque séquence dans la seconde session.

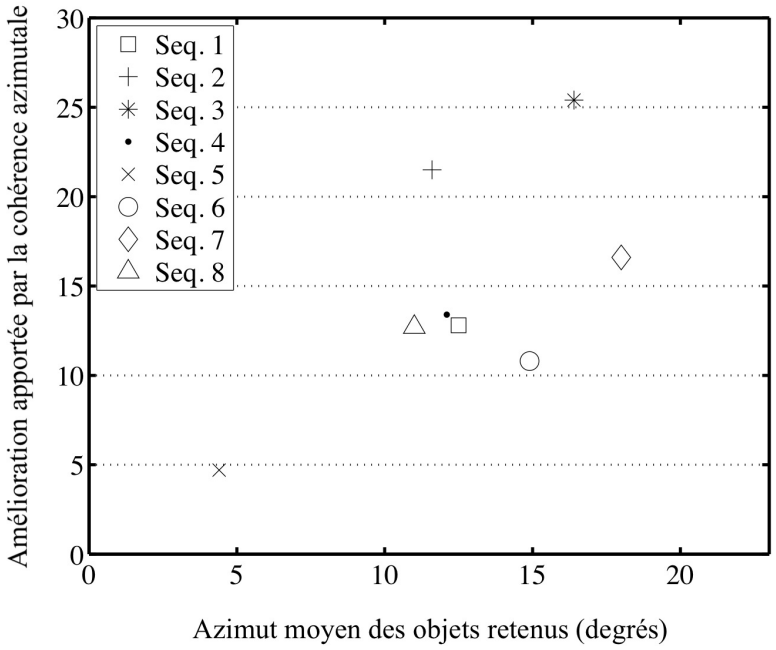


FIGURE 5.23 – Amélioration moyenne apportée par la cohérence en azimut en fonction de l'azimut moyen des objets retenus pour chaque séquence dans la seconde session

apportée par la cohérence azimutale. Certes, les séquences 2 et 3, qui ont les vitesses les plus élevées, possèdent également les améliorations les plus importantes. Cependant, on observe aussi une amélioration importante pour la séquence 7, quand bien même les objets de cette séquences ont une vitesse moyenne nulle.

Ainsi, que l'étude des corrélations se fasse sur les deux sessions ou uniquement sur la seconde session, les mêmes tendances globales émergent :

— L'amélioration apportée par la cohérence azimutale sera d'autant plus forte que la

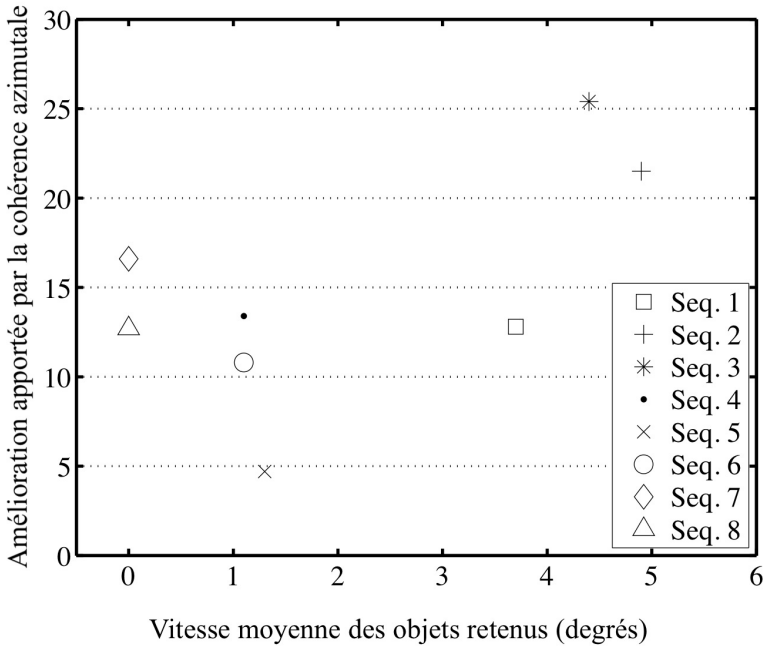


FIGURE 5.24 – Amélioration moyenne apportée par la cohérence en azimut en fonction de la vitesse moyenne des objets retenus pour chaque séquence dans la seconde session

séquence contiendra des objets présentant soit des azimuts moyens élevés, soit des vitesses moyennes élevées ;

— Cependant, cette amélioration nécessite que le son de l'objet soit suffisamment long pour pouvoir avoir une influence sur le jugement global de la séquence (comme le suggère l'exemple de l'objet « Branches » dans la séquence 5 ;

— Une vitesse moyenne élevée des objets est une condition suffisante pour que l'amélioration soit significative, mais elle n'est pas une condition nécessaire : la séquence 7, par exemple, présente une amélioration élevée quand bien même elle ne contient que des sources statiques.

Corrélations entre les caractéristiques des séquences et l'amélioration apportée par la Simulation de la Profondeur sur l'ensemble des sessions

Il n'est ici pas nécessaire de se focaliser sur la seconde session vu que l'interaction « Simulation de la profondeur × Répétition » n'est pas significative ($p = 0.337$).

La Simulation de la Profondeur n'a amélioré l'adéquation du son que pour une seule séquence : la séquence 6. Il faut donc identifier ce qui distingue la séquence 6 de toutes les autres séquences. Intuitivement, on peut se dire que la simulation de la profondeur va être cruciale pour des sources lointaines, car sinon le sujet risque de percevoir une trop grande

différence entre le rapport champ direct/champ diffus qu'on lui propose et le rapport champ direct/champ diffus qu'il rencontrerait « dans la vraie vie ».

Le fait que la séquence 6 soit significative et pas les autres ne peut pas être expliqué par des différences de divergences entre les séquences 3D-s, puisque les mêmes tendances ont été observées avec les séquences 2D (Interaction V×D non significative : p = 0.244). Une observation rapide des séquences montre qu'il n'y a que trois séquences pour lesquelles les objets sont franchement loin :

— Les 4 dernières secondes de la séquence 2, lorsque le cycliste traverse la plage de droite à gauche ;
— La séquence 3 (dialogue entre deux hommes sur une plage) ;
— La séquence 6 (un homme poussant un chariot dans un couloir), soit la séquence qui a donné lieu à une amélioration significative ;

Les deux grandes différences entre la séquence 6 et les séquences 2-3 sont que :

— la séquence 6 se déroule dans un couloir, avec des murs réfléchissants. L'image suggère donc une acoustique très présente, que l'auditeur n'entend pas dans le mixage « proximité », d'où sa « frustration ». Par contre, les séquences 2 et 3 se déroulent en extérieur, sans acoustique (si ce n'est les réflexions sur le sol et quelques rochers) : les mixages « proximité » sont donc beaucoup moins gênants pour ces séquences.
— dans la séquence 6, l'objet se déplace dans la profondeur (le chariot s'éloigne) alors que les objets lointains des séquences 2 et 3 restent à distance fixe et n'effectuent que des déplacements gauche-droite.

5.2.5 Discussion

Aucun effet du Mode Visuel

Aucun effet significatif du Mode Visuel (3D-s vs. 2D) sur les jugements des sujets n'a pu être mis en évidence, aussi bien dans la première que dans la seconde session. Ces résultats sont en accord avec ceux de Kruszielski *et al.* (2012). Les interactions impliquant le Mode Visuel ne sont pas non plus significatives, ce qui signifie que la stéréoscopie n'a pas eu d'impact sur les attentes des sujets concernant la spatialisation des objets sonores. Bien qu'en contradiction avec l'hypothèse d'André *et al.* (2012) et plusieurs témoignages d'ingénieurs du son (Gambier, 2010; Krohn, 2009), les résultats rejoignent ceux des expériences I et II, dans lesquelles nous avions déjà observé un effet limité de la stéréoscopie sur la perception des sons d'ambiance.

Effet de la Cohérence Azimutale et comparaison avec les études sur l'effet ventriloque

La cohérence audiovisuelle en azimut a significativement amélioré l'adéquation perçue du son avec l'image pour la plupart des séquences.

Le fait que la cohérence audiovisuelle en azimut ait significativement amélioré l'adéquation perçue du son est remarquable car il va à l'encontre des résultats obtenus par André *et al.* (2012), et surtout à l'encontre d'une convention cinématographique « historique » qui incite à « mixer au centre », car la latéralisation des objets serait prétendument inutile et risquerait même de gêner les spectateurs, peu habitués à une telle spatialisation (Chion, 2003). Dans notre expérience, les sujets étaient en effet peu habitués à une telle spatialisation des sources, ce qui explique peut-être pourquoi la cohérence azimutale n'a pas amélioré l'adéquation du son au début de l'expérience (session 1) : les sujets avaient besoin de temps pour s'habituer et pour « calibrer » leurs attentes. Par contre, la cohérence azimutale a bel et bien amélioré l'adéquation du son lors de la seconde session, ce qui traduit probablement une adaptation rapide des sujets à cette nouvelle façon de spatialiser les objets sonores, et nous pouvons même imaginer que les améliorations observées auraient continué à augmenter si nous avions conduit d'autres sessions supplémentaires.

Les tendances observées suggèrent donc que l'adaptation du public à la cohérence audiovisuelle en azimut ne constituerait pas un processus aussi fastidieux que le redoute Chion (2003; 2005).

Il faut tout de même remarquer que les jugements des sujets sur l'adéquation du son sont restés dans la moitié haute de l'échelle de notation (50-100), même pour les mixages « classiques », ce qui montre qu'une bande-son sans cohérence azimutale reste tout à fait satisfaisante, comme l'ont rapporté la plupart des sujets après avoir passé le test.

L'étude des corrélations entre les caractéristiques spatiales des objets et l'amélioration apportée par la cohérence azimutale a montré que cette amélioration était d'autant plus forte que la séquence contenait des objets présentant soit des azimuts moyens élevés, soit des vitesses moyennes élevées. La corrélation avec l'azimut moyen s'explique facilement : si l'azimut moyen est faible, alors cela signifie que l'objet est très proche du milieu de l'écran, et donc proche de l'enceinte centrale. Durant les séquences avec mixage « classique », les disparités angulaires entre l'objet visuel et l'objet sonore associé, diffusé sur l'enceinte centrale, étaient donc faibles et l'effet ventriloque était très efficace. Le sujet ne pouvait alors plus différencier les mixages « cohérents » des mixages « classiques ».

La corrélation positive avec la vitesse angulaire moyenne peut également s'expliquer facilement : si un objet visuel est en mouvement, alors il y a lors des mixages « classiques » non seulement des conflits de position entre objets visuel et sonore associés, mais aussi des conflits de « dynamique » (objet visuel en mouvement vs. objet sonore statique sur l'enceinte centrale). Cette remarque suggère qu'il serait intéressant de reproduire l'expérience IV sur l'effet ventriloque, mais cette fois-ci avec des objets audiovisuels en mouvement, et d'étudier comment les seuils à 50% varient en fonction de la vitesse des objets.

L'exemple de la séquence 7, dans laquelle les sources sont statiques, montre cependant que le mouvement n'est pas une condition nécessaire pour que la cohérence azimutale améliore

significativement l'adéquation du son à l'image.

Effet de la Simulation de la Profondeur

La simulation de la profondeur a amélioré l'adéquation du son à l'image dans une moindre mesure que la cohérence azimutale (+3 avec Simulation de la Profondeur contre +11 avec la Cohérence Azimutale), et n'a eu d'impact significatif que pour une seule séquence (la séquence 6, alors que la cohérence azimutale a amélioré l'adéquation du son pour 5 séquences sur 8). La non-significativité de l'interaction « Répétition × Simulation de la Profondeur » suggère que nous n'aurions pas observé avec le temps une amélioration plus importante si nous avions conduit des sessions supplémentaires.

Moulin (2015) a utilisé l'intensité et les différences binaurales créées par la WFS pour assurer une cohérence audiovisuelle en profondeur. Les améliorations de la sensation d'immersion ont été observées pour seulement 2 séquences sur 9. D'un autre côté, quand Kruszielski *et al.* (2012) ont enregistré un saxophoniste, utilisant des systèmes de prise de son et des caméras placés à différentes distances du musicien, les résultats ont montré que plus les systèmes de prise de son étaient éloignés du point de vue de la caméra, moins ils étaient jugés « adaptés ». Nous avions donc formulé l'hypothèse que des effets plus importants seraient obtenus si des indices de la perception auditive de la distance plus fondamentaux (tels que le rapport champ direct/champ diffus ou la coloration spectrale) étaient utilisés.

Cependant, l'influence de la Simulation de la Profondeur dans la présente expérience a finalement été bien plus faible que prévu. Comme il s'agit d'une séquence musicale dans l'étude de Kruszielski, il est possible que les sujets se soient davantage focalisés sur les qualités spectrales et spatiales des enregistrements par rapport à notre étude. De plus, les sujets de Kruszielski avaient une expérience du mixage et des tests subjectifs, alors que les sujets de la présente expérience étaient tous « naïfs ».

Dans le cas de la séquence 6 (un homme poussant un chariot dans un couloir), l'amélioration apportée par la Simulation de la Profondeur est très prononcée (de 35 à 66, soit presque un tiers de l'échelle de notation) et trois fois supérieure à l'amélioration apportée par la Cohérence Azimutale (de 46 à 55). La séquence 6 se différencie des autres par le fait qu'elle contient un objet lointain évoluant dans un environnement réverbérant et se déplaçant dans la profondeur. Cette observation rappelle les résultats obtenus pour la cohérence azimutale et montre que la cohérence audiovisuelle, que ce soit en azimut ou en profondeur, semble être globalement bénéfique à l'expérience du spectateur pour des sources éloignées (azimut élevé et/ou profondeur importante) et en mouvement.

5.2.6 Conclusion

Les résultats de l'expérience V ne permettent pas de valider la sous-hypothèse 2. Certes, la cohérence audiovisuelle spatiale améliore globalement l'adéquation du son à l'image, cependant cette amélioration n'est pas plus marquée en 3D-s qu'en 2D.

L'étude des corrélations entre les caractéristiques spatiales des objets et les jugements des sujets suggère que la cohérence audiovisuelle, que ce soit en azimut ou en profondeur, semble être globalement bénéfique à l'expérience du spectateur pour des sources éloignées (azimut élevé et/ou profondeur importante) et en mouvement. Ce phénomène s'explique simplement. Nous avons montré dans le chapitre 3 que l'effet ventriloque décroit lorsque l'écart entre les stimuli sonore et visuel augmente. Ainsi, le sujet dans notre expérience était plus susceptible de remarquer :

— des disparités spatiales entre son et image lors des mixages « classiques » lorsque les objets visuels étaient excentrés ;

— des disparités spatiales entre son et image lors des mixages « proximité » lorsque les objets visuels étaient distants ;

De plus, si l'objet visuel est en mouvement alors que l'objet sonore reste fixe, le spectateur risque également de détecter un conflit de "dynamiques" et de moins bien noter l'adéquation du son à l'image.

Chapitre 6

Conclusion

6.1 Influence de la stéréoscopie

Nous souhaitions vérifier dans le cadre de cette thèse l'hypothèse que la stéréoscopie modifie significativement notre perception ou nos attentes sonores par rapport à une projection en 2D (hypothèse globale). Nous avions décidé, au vu des études précédentes et des témoignages d'ingénieurs du son (résumés dans le chap. 3), de « décomposer » cette hypothèse en deux sous-hypothèses :

Sous-hypothèse 1 - La stéréoscopie change nos attentes en termes de balance frontal/surround : en 3D-s, nous souhaitons entendre « plus de surround »

Lors des expériences I et II, des séquences ont été présentées à des sujets dans leurs versions 3D-s et 2D. Dans l'expérience I, les sujets étaient dans un auditorium de mixage et devaient eux-mêmes régler la balance frontal/surround des sons d'ambiance, pour vérifier s'ils avaient une stratégie de mixage différente selon que l'image était projetée en 2D ou en 3D-s. Dans l'expérience II, les sujets étaient dans un cinéma et devaient juger des balances frontal/surround qui avaient été au préalable fixées par les expérimentateurs, afin d'étudier si les mixages 3D-s sonnaient plus frontaux, plus « surround », ou alors de la même façon que les mixages 2D.

Bien que ces deux expériences aient été menées dans des lieux différents, avec des tâches et des séquences différentes, elles ont toutes les deux montré que **l'influence de la stéréoscopie était limitée et n'apparaissait que pour quelques séquences**. L'expérience II suggère également que cette influence devient **fragile au cours du temps** et **ne concerne pas toutes les places** dans la salle de cinéma.

Une troisième expérience a été conduite pour vérifier si les séquences qui avait été significativement impactées par la stéréoscopie étaient celles dont les différences entre versions 2D et 3D-s étaient les plus importantes en termes de profondeur visuelle perçue. Cependant, aucune corrélation substantielle n'a pu être observée. Par contre, une bonne corrélation a pu

être observée entre les tailles des boîtes scéniques des séquences et les différences de balances perçues au « Sweet Spot » dans l'expérience II : plus la boite scénique était grande, plus les différences de balances perçues entre versions 2D et 3D-s étaient importantes. Les différences de perception sonore entre versions 2D et 3D-s d'une séquence semblent donc bien dépendre directement de la « quantité » de stéréoscopie présente dans la version 3D-s.

Sous-hypothèse 2 - La stéréoscopie change nos attentes en termes de spatialisation des objets sonores : en 3D-s, une plus grande cohérence spatiale entre le son et l'image est attendue.

L'expérience IV ne portait pas spécifiquement sur l'influence de la stéréoscopie : des images 3D-s d'un personnage en train de parler étaient présentées à des sujets droit devant eux. La voix de l'homme pouvait être reproduite sur différentes enceintes, qui créaient des disparités plus ou moins grandes entre le son et l'image. Pour chaque présentation, les sujets devaient indiquer si la voix semblait provenir ou non de la même direction que la bouche du personnage. Les « seuils à 50% » ont été mesurés pour des sources sonores variant à la fois en azimut et en élévation par rapport à la source visuelle, afin que l'effet ventriloque dans le plan vertical et l'effet ventriloque dans le plan horizontal puissent être directement comparés. Les résultats ont montré que l'effet ventriloque pouvait fonctionner en élévation à des écarts angulaires très élevés (certains sujets continuaient de percevoir la voix du personnage sur sa bouche même lorsque le son était diffusé dans leur dos), ce qui suggérait que **l'influence de la cohérence audiovisuelle en élévation sur la qualité d'expérience audiovisuelle était négligeable dès lors qu'un mixage était déjà cohérent en azimut**. Nous avons donc décidé de ne pas retenir la dimension verticale pour la suite de notre étude et de nous concentrer sur l'azimut et la profondeur.

Dans l'expérience V, les sujets devaient évaluer à quel point les bandes-son qui leur étaient proposées étaient « adaptées » à l'image pour 8 séquences projetées en 2D et en 3D-s :

— les sources sonores (dialogues et effets *in*) pouvaient être soit diffusées sur l'enceinte centrale (comme il est d'usage dans les productions professionnelles) soit au même azimut que leur correspondant visuel (cohérence azimutale) ;

— dans certaines bandes-son, les enregistrements originaux en proximité des sources sonores étaient utilisés, alors que des traitements numériques (équalisation, niveau, réverbération) et des prises de son en champ diffus étaient utilisés dans d'autres bandes-son, afin de mieux respecter la position en profondeur des sources visuelles associées (simulation de la profondeur).

Les résultats ont montré que la cohérence azimutale pouvait efficacement améliorer l'adéquation du son à l'image, surtout dans le cas d'objets en mouvement et loin du milieu de l'écran. En profondeur, une amélioration a également pu être constatée, mais uniquement

pour une séquence. Ainsi, **les sujets préféraient globalement les mixages cohérents**, mais cette préférence était **indépendante de la stéréoscopie**, puisqu'**aucun effet du mode visuel (3D-s vs. 2D) sur les jugements des sujets** n'a pu être observé.

Les résultats de nos expériences s'accordent donc sur le fait que **l'influence de la stéréoscopie sur la perception du son au cinéma est faible**. Il est cependant à noter que les séquences utilisées dans le cadre de notre étude provenaient uniquement de productions à petit ou moyen budget, et non de *blockbusters* tels que *Gravity* ou *Avatar*, qui constituent la majorité des films tournés en 3D-s et dont le contenu joue beaucoup plus sur le « spectaculaire » que nos séquences, avec scènes d'action, effets de jaillissement et parfois images de synthèse. Nous n'avons pas non plus traité le cas d'objets hors-champ et d'objets entrant ou sortant du cadre. Il est donc clair que nous n'avons pas exploré la totalité des possibilités qu'offre la 3D-s, et que nos résultats ne sauraient être généralisés à tout type de contenus 3D-s.

6.2 Confrontation des nouveaux systèmes de spatialisation aux résultats de la présente étude

Dans le chapitre 1.2.5, nous avons vu que les améliorations proposées par les nouveaux systèmes de spatialisation pouvaient se résumer en deux points :

— une plus grande égalité entre enceintes frontales et enceintes “surround" ;

— une plus grande cohérence audiovisuelle spatiale.

Dans cette thèse, nous avons vu que :

— en effet, les sujets souhaitaient parfois entendre plus de surround en 3D-s, mais cela concernait un nombre très limité de séquences ;

— concernant la spatialisation des objets sonores, une plus grande cohérence audiovisuelle semble en effet souhaitable en azimut ;

— les résultats de l'expérience V suggèrent que la cohérence audiovisuelle n'est pas nécessaire dès lors que l'effet ventriloque fonctionne en configuration « classique » (c'est-à-dire lorsque les objets sonores sont diffusés sur l'enceinte centrale). Ce constat conforte *a posteriori* notre hypothèse que la cohérence audiovisuelle en élévation est superflue. Nous ne pouvons pas conclure pour autant que l'ajout d'enceintes zénithales est inutile au cinéma : ces enceintes trouvent peut-être leur justification dans d'autres contextes. Dupas (2007), par exemple, a montré que la diffusion de sons d'ambiance zénithaux pouvait accroître la sensation d'immersion des sujets. De plus, nous avons vu dans l'expérience IV que l'effet ventriloque était très efficace en élévation pour une source statique, mais il faudrait vérifier si l'effet fonctionne toujours aussi bien pour une source en mouvement.

Troisième partie

Annexes

Annexe A

Publications liées à la thèse

A.1 Revues

E. Hendrickx, M. Paquier, and V. Koehl (2014). The Influence of Stereoscopy on the Sound Mixing of Movies : A Study on the Front/Rear Balance of Ambience. *Journal of the Audio Engineering Society*, 62(11), 723-735.

URL : http ://www.aes.org/e-lib/browse.cfm ?elib=17548

E. Hendrickx, M. Paquier, and V. Koehl (2015). Audiovisual Spatial Coherence for 2D and Stereoscopic-3D movies. *Journal of the Audio Engineering Society*, 63(11), 889-899.

URL : http ://www.aes.org/e-lib/browse.cfm ?elib=18049

E. Hendrickx, M. Paquier, and V. Koehl (2015). Ventriloquism effect with sound stimuli varying in both azimuth and elevation. *Journal of the Acoustical Society of America*, 138(6), 3686-3697.

URL : http ://scitation.aip.org/content/asa/journal/jasa/138/6/10.1121/1.4937758

A.2 Conférences

E. Hendrickx, M. Paquier, and V. Koehl. Does Stereoscopy Change Our Perception of Soundtracks ? *Proceedings of the 57th AES International Conference : The Future of Audio Entertainment Technology-Cinema, Television and the Internet* (2015).

URL : http ://www.aes.org/e-lib/browse.cfm ?elib=17600

E. Hendrickx, M. Paquier, and V. Koehl. Influence de la stéréoscopie sur le mixage des ambiances « surround » au cinéma. *Actes du 12e Congrès Français d'Acoustique* (2014) : 1771-1777.

URL : https ://hal.archives-ouvertes.fr/hal-00986896/

E. Hendrickx, M. Paquier, and V. Koehl. Should a movie have two different soundtracks for its stereoscopic and non-stereoscopic versions ? A case study on the front/rear balance. *Proceedings of the International Conference on 3D Imaging* (2013).

Prix 2013 de la meilleure recherche scientifique lors de la *Conférence Internationale sur l'Imagerie 3D* (Lièges, Belgique).

URL : http ://ieeexplore.ieee.org/stamp/stamp.jsp ?tp=&arnumber=6732079&isnumber=6732069

Annexe B

Systèmes d'enregistrement utilisés dans l'expérience I

3 séquences ont été enregistrées avec un système double-M/S (Wittek *et al.*, 2006). Ce système, présenté dans la Fig. B.1, est composé de deux microphones cardioïdes (un microphone pointant vers l'avant, l'autre vers l'arrière) et d'un microphone bi-directionnel (à 90° par rapport à l'axe des deux autres microphones).

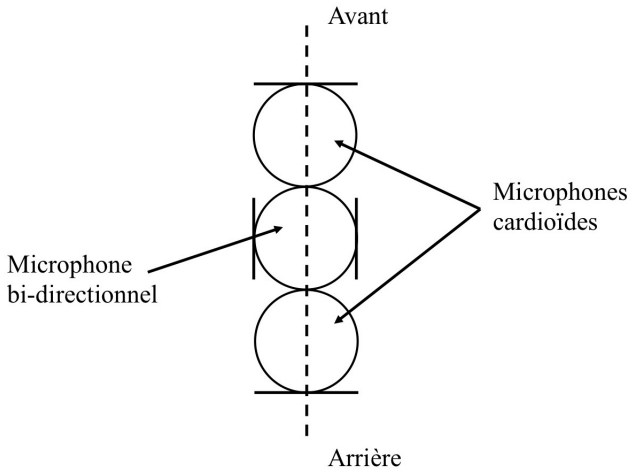


FIGURE B.1 – Système double-M/S. Les trois microphones sont supposés coïncidents.

2 séquences ont été enregistrées avec un système Double-ORTF (Czyzewski *et al.*, 2002). Ce système, présenté dans la Fig. B.2, est composé de 4 microphones cardioïdes formant un carré, avec un angle de 110° entre les deux microphones frontaux et entre les deux microphones arrières.

2 séquences ont été enregistrées avec un système Fukada Tree (Hiekkanen *et al.*, 2007). Il existe plusieurs configurations possibles pour cet arbre multicanal. Nous avons choisi celle utilisée par Hiekkanen *et al.* (2007) (présentée dans la Fig. B.3), qui est composée de trois microphones cardioïdes en triangle pour les canaux frontaux, et deux microphones cardioïdes pointant vers l'arrière pour les canaux « surround ».

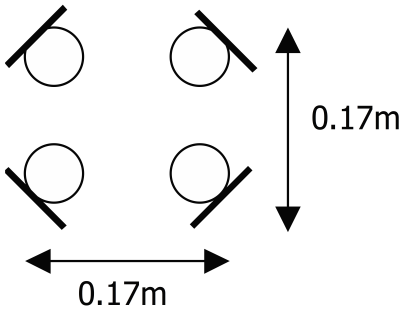


FIGURE B.2 – Système double-ORTF. D'après Czyzewski *et al.* (2002).

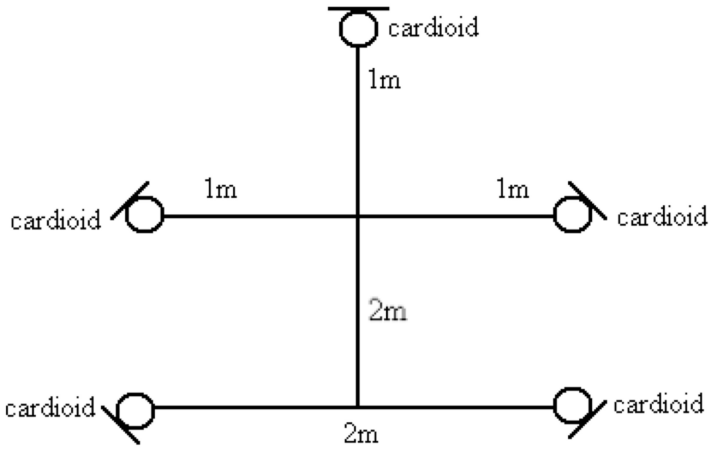


FIGURE B.3 – Système Fukada. D'après Hiekkanen *et al.* (2007).

Annexe C

Récapitulatif détaillé des séquences utilisées dans l'expérience I

Séquence 1

Durée : 35 secondes. Intérieur.

Travelling avant très lent se rapprochant progressivement de la violoncelliste : d'un plan large à un plan serré sur le violoncelle.

Dans cette séquence, une violoncelliste joue la 3ème suite pour violoncelle de Bach, dans une église très réverbérante. La séquence a été enregistrée à l'aide d'un système Fukada Tree (voir Annexes).

Le tableau ci-dessous présente les niveaux sonores de chaque piste audio correspondant à la balance frontal/surround « nominale » fixée par les expérimentateurs, et qui servira de référence pour les analyses à venir.

Piste audio	Niveau sonore moyen (dB RMS)
C	Pas de son
L (Réverbération et son direct)	-29.0
R (Réverbération et son direct)	-26.8
Ls (Réverbération)	-34.5
Rs (Réverbération)	-34.9

TABLEAU C.1 – Niveau sonore moyen de chaque piste audio pour la séquence 1 de l'expérience I.

FIGURE C.1 – Evolution du travelling dans la séquence 1 de l'expérience I. Captures d'image.

Séquence 2

Durée : 29 secondes.

Plan fixe et serré sur la violoncelliste. Intérieur.

Dans cette séquence, la violoncelliste de la séquence 1 s'adresse à la caméra pour présenter la pièce qu'elle vient de jouer.

La séquence a été enregistrée à l'aide d'un système Fukada Tree (voir Annexes).

Le tableau ci-dessous présente les niveaux sonores de chaque piste audio.

Piste audio	Niveau sonore moyen (dB RMS)
C	Pas de son
L (Réverbération et son direct)	-42.2
R (Réverbération et son direct)	-40.3
Ls (Réverbération)	-50.5
Rs (Réverbération)	-51.8

Tableau C.2 – Niveau sonore moyen de chaque piste audio pour la séquence 2 de l'expérience I.

Figure C.2 – Séquence 2 de l'expérience I. Capture d'image.

Séquence 3

Durée : 30 secondes.

Plan large et fixe. Intérieur.

Cette séquence montre l'intérieur d'une crèche, vide.

Les ambiances, captées sur place en stéréophonie décorrélée, sont constituées de bruits venant de l'extérieur d'enfants en train de jouer dans la cour.

Le tableau ci-dessous présente les niveaux sonores de chaque piste audio.

Piste audio	Niveau sonore moyen (dB RMS)
C	Pas de son
L (Ambiance crèche)	-51.2
R (Ambiance crèche)	-51.4
Ls (Ambiance crèche)	-58.4
Rs (Ambiance crèche)	-57.2

Tableau C.3 – Niveau sonore moyen de chaque piste audio pour la séquence 3 de l'expérience I.

Figure C.3 – Séquence 3 de l'expérience I. Capture d'image.

Séquence 4

Durée : 23 secondes.

Plan fixe et serré sur les deux comédiennes. Intérieur.

Dans cette séquence, deux femmes dialoguent dans un café.

Les ambiances ont été captées sur place en stéréophonie décorrélée. On peut y entendre des gens en train de discuter, des bruits de tasses posées sur un bar, une porte qui s'ouvre et se referme et une petite musique d'ambiance. Les dialogues ont été enregistrés avec une perche et un microphone canon (Neumann KMR 81), puis diffusés sur l'enceinte centrale.

Le tableau ci-dessous présente les niveaux sonores de chaque piste audio.

Piste audio	Niveau sonore moyen (dB RMS)
C (Dialogues)	-35.8
L (Ambiance café)	-50.0
R (Ambiance café)	-49.0
Ls (Ambiance café)	-56.9
Rs (Ambiance café)	-55.5

TABLEAU C.4 – Niveau sonore moyen de chaque piste audio pour la séquence 4 de l'expérience I.

FIGURE C.4 – Séquence 4 de l'expérience I. Capture d'image.

Séquence 5

Durée : 23 secondes.

Il s'agit de la même séquence que la séquence 4, sauf que les dialogues, toujours diffusés dans l'enceinte centrale, sont diffusés 6 dB plus fort.

Le tableau ci-dessous présente les niveaux sonores de chaque piste audio.

Piste audio	Niveau sonore moyen (dB RMS)
C (Dialogues)	-29.8
L (Ambiance café)	-50.0
R (Ambiance café)	-49.0
Ls (Ambiance café)	-56.9
Rs (Ambiance café)	-55.5

TABLEAU C.5 – Niveau sonore moyen de chaque piste audio pour la séquence 5 de l'expérience I.

FIGURE C.5 – Séquence 5 de l'expérience I. Capture d'image.

Séquence 6

Durée : 30 secondes.

Caméra épaule, suivant les deux comédiennes. Plan serré. Extérieur.

Dans cette séquence, deux femmes marchent dans une forêt.

Les ambiances ont été captées sur place en stéréophonie décorrélée. On peut y entendre une rumeur de ville au loin et des oiseaux. Les bruits de pas ont été enregistrés avec une perche et un microphone canon (Neumann KMR 81), puis diffusés sur l'enceinte centrale.

Le tableau ci-dessous présente les niveaux sonores de chaque piste audio.

Piste audio	Niveau sonore moyen (dB RMS)
C (Bruits de pas)	-53.0
L (Ambiance forêt)	-47.2
R (Ambiance forêt)	-48.6
Ls (Ambiance forêt)	-48.6
Rs (Ambiance forêt)	-48.1

TABLEAU C.6 – Niveau sonore moyen de chaque piste audio pour la séquence 6 de l'expérience I.

FIGURE C.6 – Séquence 6 de l'expérience I. Capture d'image.

Séquence 7

Durée : 30 secondes.

Plan fixe et large. Extérieur.

Cette séquence montre un pont, traversé par de nombreuses voitures puis un tram en jaillissement, qui sort de l'image à droite de l'écran.

Les ambiances ont été captées sur place avec un système double ORTF et sont synchrones à l'image. On peut y entendre une rumeur de ville, avec le bruit des voitures et du tram.

Le tableau ci-dessous présente les niveaux sonores de chaque piste audio.

Piste audio	Niveau sonore moyen (dB RMS)
C	Pas de son
L (Ambiance ville)	-29.0
R (Ambiance ville)	-27.4
Ls (Ambiance ville)	-27.6
Rs (Ambiance ville)	-26.3

TABLEAU C.7 – Niveau sonore moyen de chaque piste audio pour la séquence 7 de l'expérience I.

FIGURE C.7 – Séquence 7 de l'expérience I. Captures d'image.

Séquence 8

Durée : 30 secondes.

Plan fixe et large. Extérieur.

Dans cette séquence, on aperçoit un port avec un bateau au premier plan, et d'autres bateaux dans la profondeur.

Les ambiances ont été captées sur place avec un système double M/S. On peut y entendre des bruits de vent, de cliquetis de poulies, de clapotis d'eau contre les barques, quelques voix au loin, ainsi qu'un hélicoptère très lointain.

Le tableau ci-dessous présente les niveaux sonores de chaque piste audio.

Piste audio	Niveau sonore moyen (dB RMS)
C	Pas de son
L (Ambiance port)	-43.7
R (Ambiance port)	-44.3
Ls (Ambiance port)	-42.4
Rs (Ambiance port)	-43.5

Tableau C.8 – Niveau sonore moyen de chaque piste audio pour la séquence 8 de l'expérience I.

Figure C.8 – Séquence 8 de l'expérience I. Capture d'image.

Séquence 9

Durée : 29 secondes.

Plan fixe et large. Extérieur.

Cette séquence montre une mer calme, vue de la proue d'un bateau.

Les ambiances ont été captées sur place avec un système Double-M/S. On peut y entendre une ambiance de mer, avec le cliquetis des poulies du bateau et quelques claquements de la voile sous l'effet du vent.

Le tableau ci-dessous présente les niveaux sonores de chaque piste audio.

Piste audio	Niveau sonore moyen (dB RMS)
C	Pas de son
L (Ambiance mer)	-39.8
R (Ambiance mer)	-40.3
Ls (Ambiance mer)	-41.9
Rs (Ambiance mer)	-42.6

TABLEAU C.9 – Niveau sonore moyen de chaque piste audio pour la séquence 9 de l'expérience I.

FIGURE C.9 – Séquence 9 de l'expérience I. Capture d'image.

Séquence 10

Durée : 23 secondes.

Plan fixe et moyen. Extérieur.

Dans cette séquence, deux marins à la proue d'un bateau effectuent des manoeuvres.

Les ambiances ont été captées sur place avec un système double-M/S et sont synchrones à l'image. On peut y entendre du vent, le cliquetis des poulies du bateau, la voix des deux marins et le bruit des vagues.

Le tableau ci-dessous présente les niveaux sonores de chaque piste audio.

Piste audio	Niveau sonore moyen (dB RMS)
C	Pas de son
L (Ambiance mer)	-36.0
R (Ambiance mer)	-36.1
Ls (Ambiance mer)	-41.9
Rs (Ambiance mer)	-41.9

TABLEAU C.10 – Niveau sonore moyen de chaque piste audio pour la séquence 10 de l'expérience I.

FIGURE C.10 – Séquence 10 de l'expérience I. Capture d'image.

Séquence 11

Durée : 30 secondes.

Plan fixe et large. Extérieur.

Cette séquence montre une foule dense dans une rue lors d'un grand évènement.

Les ambiances, captées sur place en Double-ORTF, sont principalement constituées d'une rumeur de foule.

Le tableau ci-dessous présente les niveaux sonores de chaque piste audio.

Piste audio	Niveau sonore moyen (dB RMS)
C	Pas de son
L (Ambiance foule)	-40.9
R (Ambiance foule)	-41.4
Ls (Ambiance foule)	-41.0
Rs (Ambiance foule)	-41.4

Tableau C.11 – Niveau sonore moyen de chaque piste audio pour la séquence 11 de l'expérience I.

Figure C.11 – Séquence 11 de l'expérience I. Capture d'image.

Annexe D

Exploration des résultats de l'expérience I

Les Fig. D.1 et D.2 montrent les diagrammes en boîte obtenus avec images 2D et images 3D-s pour la variable ΔG, séquence par séquence, pour les deux sessions.

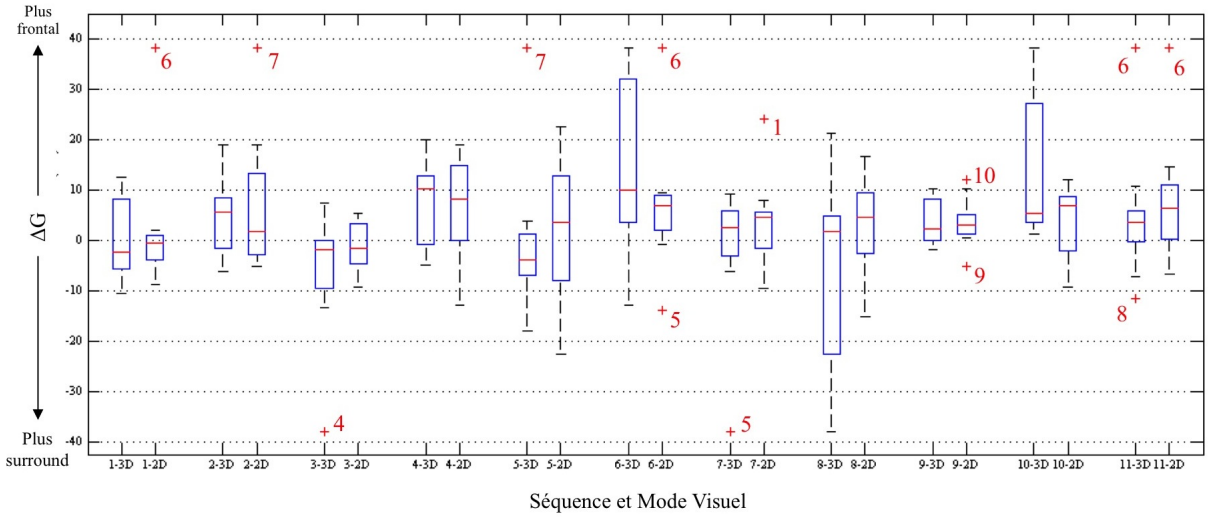


FIGURE D.1 – Diagrammes en boîte obtenus pour ΔG lors de la session 1 pour chaque séquence. Une croix rouge indique un « outlier ». Le chiffre rouge à coté indique le numéro du sujet concerné.

Nous constatons que les distributions sont souvent centrées sur une valeur proche de $\Delta G = 0$, ce qui montre que les sujets ne se sont globalement pas écartés substantiellement du mixage « nominal » proposé par les expérimentateurs. Pour beaucoup de distributions, les « moustaches » ne sont pas de la même taille et les boîtes ne sont pas centrées sur la médiane, ce qui suggère des distributions asymétriques et donc non normales.

Nous observons également un nombre important d'« outliers », ce qui suggère qu'il faudra peut-être envisager des tests non-paramétriques pour l'analyse des données, moins sensibles aux « outliers » que les tests paramétriques. Ces « outliers » correspondent soit à des mixages très frontaux (ce qui en soi reste concevable, puisque nous sommes, avec la stéréophonie,

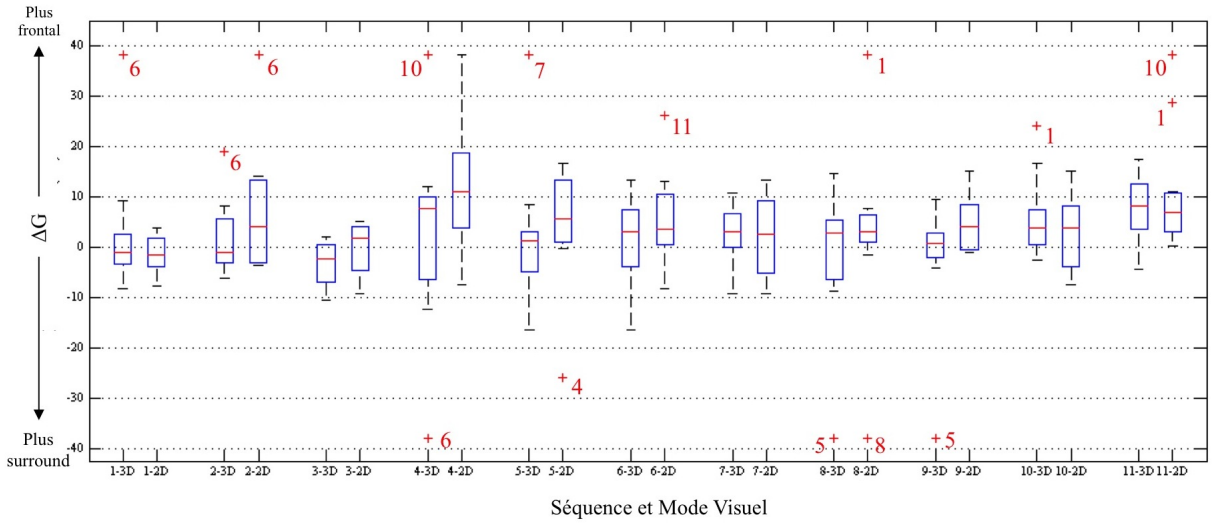


FIGURE D.2 – Diagrammes en boîte obtenus pour ΔG lors de la session 2 pour chaque séquence. Une croix rouge indique un « outlier ». Le chiffre rouge à coté indique le numéro du sujet concerné.

culturellement imprégnés de reproductions sonores purement frontales), soit, plus bizarrement, à des mixages avec du son presque uniquement dans les enceintes « surround ». Une analyse plus en détail montre qu'aucun sujet n'a été un « outlier » systématique, sauf le sujet 6 qui sur les 22 conditions a été à 8 reprises très éloigné du reste des observations.

Nous n'observons pas à première vue sur les graphes de différence flagrante entre mixages avec images 3D-s et mixages avec images 2D, sauf pour quelques cas : les séquences 4 et 5 dans la session 2, par exemple, semblent avoir été « mixées plus frontales » en 2D qu'en 3D-s.

Nous remarquons dans la session 1 trois conditions pour lesquelles les boîtes sont très larges : séquence 6 en 3D-s, séquence 8 en 3D-s et séquence 10 en 3D-s. Ces boîtes sont beaucoup plus resserrées dans la session 2, ce qui montre que les mixages ont gagné en homogénéité d'une session à l'autre. Cependant, ce resserrement des distributions n'est pas forcément observé pour les autres conditions : on ne peut donc pas dire qu'on assiste à une homogénéisation globale des mixages d'une session à l'autre.

Annexe E

Récapitulatif détaillé des séquences utilisées dans les expériences II et III

Séquence 1

Durée : 23 secondes.

Plan fixe et serré sur les deux comédiennes. Intérieur.

Il s'agit de la séquence 4 de l'expérience I. Dans cette séquence, deux femmes dialoguent dans un café.

Les ambiances ont été captées sur place en stéréophonie décorrélée. On peut y entendre des gens en train de discuter, des bruits de tasses posées sur un bar, une porte qui s'ouvre et se referme et une petite musique d'ambiance. Les dialogues ont été enregistrés avec une perche et un microphone canon (Neumann KMR 81), puis diffusés sur l'enceinte centrale.

Les tableaux ci-dessous présentent les niveaux sonores de chaque piste audio pour le mixage A et le mixage B.

Piste audio	Niveau sonore moyen (dB RMS)
C (dialogues)	-35.7
L (ambiance café)	-50.0
R (ambiance café)	-49.0
Ls (ambiance café)	-59.1
Rs (ambiance café)	-57.4

Tableau E.1 – Niveau sonore moyen de chaque piste audio pour le mixage A de la séquence 1 de l'expérience II.

Piste audio	Niveau sonore moyen (dB RMS)
C (dialogues)	-35.7
L (ambiance café)	-51.6
R (ambiance café)	-50.6
Ls (ambiance café)	-53.8
Rs (ambiance café)	-52.4

Tableau E.2 – Niveau sonore moyen de chaque piste audio pour le mixage B de la séquence 1 de l'expérience II.

Figure E.1 – Séquence 1 de l'expérience II. Capture d'image.

Séquence 2

Durée : 23 secondes.

Plan fixe et moyen. Extérieur.

Il s'agit de la séquence 10 de l'expérience I. Dans cette séquence, deux marins à la proue d'un bateau effectuent des manoeuvres.

Les ambiances ont été captées sur place avec un système double-M/S et sont synchrones à l'image. On peut y entendre du vent, le cliquetis des poulies du bateau et le bruit des vagues.

Les tableaux ci-dessous présentent les niveaux sonores de chaque piste audio pour le mixage A et le mixage B.

Piste audio	Niveau sonore moyen (dB RMS)
C	Pas de son
L (ambiance mer)	-32.1
R (ambiance mer)	-32.2
Ls (ambiance mer)	-45.0
Rs (ambiance mer)	-45.0

Tableau E.3 – Niveau sonore moyen de chaque piste audio pour le mixage A de la séquence 2 de l'expérience II.

Piste audio	Niveau sonore moyen (dB RMS)
C	Pas de son
L (ambiance mer)	-34.9
R (ambiance mer)	-35.0
Ls (ambiance mer)	-37.8
Rs (ambiance mer)	-37.8

Tableau E.4 – Niveau sonore moyen de chaque piste audio pour le mixage B de la séquence 2 de l'expérience II.

Figure E.2 – Séquence 2 de l'expérience II. Capture d'image.

Séquence 3

Durée : 20 secondes.

Plan fixe et large. Extérieur.

Il s'agit de la séquence 11 de l'expérience I. Cette séquence montre une foule dense dans une rue lors d'un grand évènement.

Les ambiances, captées sur place en Double-ORTF, sont principalement constituées d'une rumeur de foule.

Les tableaux ci-dessous présentent les niveaux sonores de chaque piste audio pour le mixage A et le mixage B.

Piste audio	Niveau sonore moyen (dB RMS)
C	Pas de son
L (ambiance foule)	-36.4
R (ambiance foule)	-36.8
Ls (ambiance foule)	-50.1
Rs (ambiance foule)	-50.4

Tableau E.5 – Niveau sonore moyen de chaque piste audio pour le mixage A de la séquence 3 de l'expérience II.

Piste audio	Niveau sonore moyen (dB RMS)
C	Pas de son
L (ambiance foule)	-39.0
R (ambiance foule)	-39.4
Ls (ambiance foule)	-39.4
Rs (ambiance foule)	-39.6

Tableau E.6 – Niveau sonore moyen de chaque piste audio pour le mixage B de la séquence 3 de l'expérience II.

Figure E.3 – Séquence 3 de l'expérience II. Capture d'image.

Séquence 4

Durée : 20 secondes.

Il s'agit d'un plan moyen sur deux personnages. La caméra, sur grue, pivote lentement autour des acteurs. Extérieur.

Dans cette séquence, un enfant et un homme sont assis dans une arbre et discutent.

Les ambiances, captées sur place en stéréophonie décorrélée, sont constituées d'un léger vent avec des oiseaux. Les dialogues sont reproduits sur l'enceinte centrale.

Les tableaux ci-dessous présentent les niveaux sonores de chaque piste audio pour le mixage A et le mixage B.

Piste audio	Niveau sonore moyen (dB RMS)
C (dialogues)	-43.3
L (ambiance forêt)	-62.2
R (ambiance forêt)	-62.0
Ls (ambiance forêt)	-69.9
Rs (ambiance forêt)	-70.5

Tableau E.7 – Niveau sonore moyen de chaque piste audio pour le mixage A de la séquence 4 de l'expérience II.

Piste audio	Niveau sonore moyen (dB RMS)
C (dialogues)	-43.7
L (ambiance forêt)	-63.7
R (ambiance forêt)	-63.3
Ls (ambiance forêt)	-60.4
Rs (ambiance forêt)	-61.1

Tableau E.8 – Niveau sonore moyen de chaque piste audio pour le mixage B de la séquence 4 de l'expérience II.

FIGURE E.4 – Séquence 4 de l'expérience II. Captures d'image.

Séquence 5

Durée : 23 secondes.

Champ-Contrechamp en plans serrés.

Dans cette séquence, un homme regarde deux enfants jouer sous une gigantesque nappe.

Les ambiances, captées sur place en stéréophonie décorrélée, sont constituées des bruits de la nappe et de bruits de vent.

Les tableaux ci-dessous présentent les niveaux sonores de chaque piste audio pour le mixage A et le mixage B.

Piste audio	Niveau sonore moyen (dB RMS)
C (rires des enfants)	-40.0
L (ambiance nappe)	-44.4
R (ambiance nappe)	-44.2
Ls (ambiance nappe)	-52.8
Rs (ambiance nappe)	-53.5

TABLEAU E.9 – Niveau sonore moyen de chaque piste audio pour le mixage A de la séquence 5 de l'expérience II.

Piste audio	Niveau sonore moyen (dB RMS)
C (rires des enfants)	-41.9
L (ambiance nappe)	-45.8
R (ambiance nappe)	-45.6
Ls (ambiance nappe)	-43.9
Rs (ambiance nappe)	-44.2

TABLEAU E.10 – Niveau sonore moyen de chaque piste audio pour le mixage B de la séquence 5 de l'expérience II.

FIGURE E.5 – Séquence 5 de l'expérience II. Captures d'image.

Séquence 6

Durée : 21 secondes.

Lent travelling arrière, serré sur les comédiennes. Intérieur.

Trois jeunes filles dansent dans une chapelle.

Sur le canal central se trouvent principalement les bruits de respiration et de pas des trois danseuses. Dans les autre canaux, on entend principalement la musique. Les tableaux ci-dessous présentent les niveaux sonores de chaque piste audio pour le mixage A et le mixage B.

Piste audio	Niveau sonore moyen (dB RMS)
C (danseuses)	-33.6
L (musique)	-30.3
R (musique)	-30.7
Ls (musique)	-35.8
Rs (musique)	-37.2

Tableau E.11 – Niveau sonore moyen de chaque piste audio pour le mixage A de la séquence 6 de l'expérience II.

Piste audio	Niveau sonore moyen (dB RMS)
C (danseuses)	-34.6
L (musique)	-31.3
R (musique)	-31.7
Ls (musique)	-31.0
Rs (musique)	-32.0

Tableau E.12 – Niveau sonore moyen de chaque piste audio pour le mixage B de la séquence 6 de l'expérience II.

Figure E.6 – Séquence 6 de l'expérience II. Capture d'image.

Séquence 7

Durée : 20 secondes.

Travelling avant très lent suivant une comédienne. Intérieur. Dans cette séquence, une jeune fille marche silencieusement dans une chapelle.

Les ambiances, captées sur place en stéréophonie décorrélée, sont principalement constituées d'un fond d'air pesant, avec léger vent, grillons et oiseaux lointains. Les bruits de pas de la jeune fille ont été enregistrés en monophonie et sont reproduits sur l'enceinte centrale.

Les tableaux ci-dessous présentent les niveaux sonores de chaque piste audio pour le mixage A et le mixage B.

Piste audio	Niveau sonore moyen (dB RMS)
C (bruits de pas)	-86.1
L (ambiance chapelle)	-53.8
R (ambiance chapelle)	-54.0
Ls (ambiance chapelle)	-67.9
Rs (ambiance chapelle)	-69.8

Tableau E.13 – Niveau sonore moyen de chaque piste audio pour le mixage A de la séquence 7 de l'expérience II.

Piste audio	Niveau sonore moyen (dB RMS)
C (bruits de pas)	-87.2
L (ambiance chapelle)	-54.8
R (ambiance chapelle)	-55.1
Ls (ambiance chapelle)	-60.9
Rs (ambiance chapelle)	-62.9

Tableau E.14 – Niveau sonore moyen de chaque piste audio pour le mixage B de la séquence 7 de l'expérience II.

Figure E.7 – Séquence 7 de l'expérience II. Capture d'image.

Séquence 8

Durée : 21 secondes.

Il s'agit d'une scène d'action, montée très rapidement (plans courts, nombreux « cut ») avec alternance de plans larges, moyens et serrés. Extérieur.

Cette séquence montre un match de rugby particulièrement intense.

Les ambiances, captées lors du tournage en Double-M/S, sont principalement constituées de bruits de foules et des bruits de pas et grognements des joueurs. Le canal central est constitué de prises de son monophoniques, avec principalement les voix des joueurs.

Les tableaux ci-dessous présentent les niveaux sonores de chaque piste audio pour le mixage A et le mixage B.

Piste audio	Niveau sonore moyen (dB RMS)
C (voix des joueurs)	-22.8
L (ambiance match)	-21.8
R (ambiance match)	-19.9
Ls (ambiance match)	-39.9
Rs (ambiance match)	-38.6

Tableau E.15 – Niveau sonore moyen de chaque piste audio pour le mixage A de la séquence 8 de l'expérience II.

Piste audio	Niveau sonore moyen (dB RMS)
C	-24.7
L (ambiance match)	-23.8
R (ambiance match)	-21.9
Ls (ambiance match)	-29.4
Rs (ambiance match)	-28.2

Tableau E.16 – Niveau sonore moyen de chaque piste audio pour le mixage B de la séquence 8 de l'expérience II.

FIGURE E.8 – Séquence 8 de l'expérience II. Captures d'image.

Annexe F

Exploration des résultats de l'expérience II

Influence du Mode Visuel : 3D-s vs. 2D

La Fig. F.1 compare les distributions obtenues pour les stimuli 3D-s et 2D, toutes autres variables indépendantes confondues.

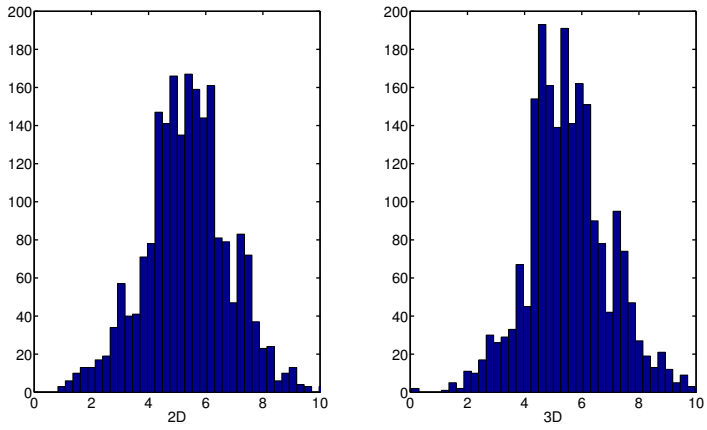


FIGURE F.1 – Comparaison des distributions obtenues pour le Mode Visuel 3D-s et 2D, toutes autres variables indépendantes confondues.

L'observation des histogrammes montre que :

— les distributions sont symétriques et centrées sur le milieu de l'échelle. Nous pouvons donc supposer que les sujets ont globalement trouvé les mixages plutôt équilibrés, aussi bien avec l'image en 2D qu'avec l'image en 3D-s ;

— il n'y a pas à première vue de différence flagrante entre 2D et 3D-s. La stéréoscopie ne semble donc pas avoir substantiellement influencé la perception de la balance frontal/surround ;

— les distributions semblent multi-modales. Les différents modes peuvent être dus aux séquences, à des groupes de sujets, aux mixages A et B ou à un effet de la session.

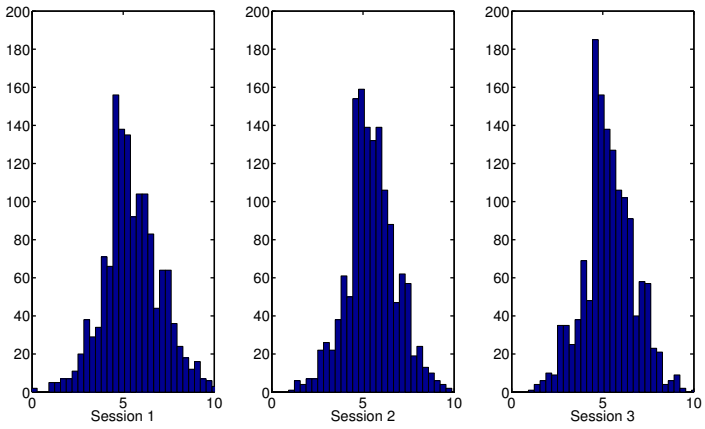


FIGURE F.2 – Comparaison des distributions obtenues pour les trois sessions, toutes autres variables indépendantes confondues.

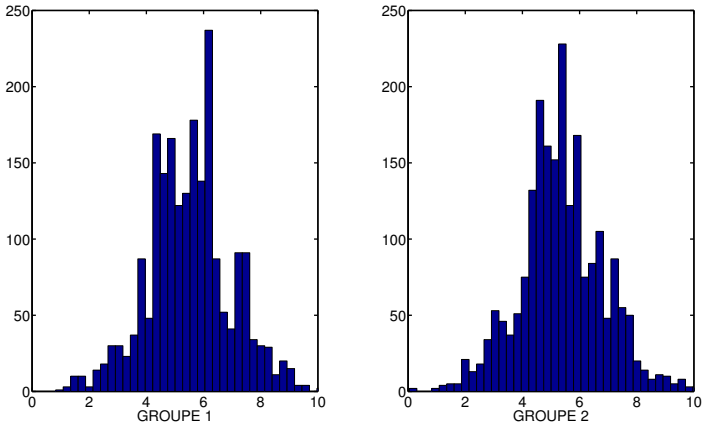


FIGURE F.3 – Comparaison des distributions obtenues pour le groupe 1 en proximité de l'écran et pour le groupe 2 au « Sweet Spot », toutes autres variables indépendantes confondues.

Influence de la Répétition

La Fig. F.2 compare les distributions obtenues pour les trois sessions du test.

L'observation des histogrammes montre que :

— nous retrouvons la tendance globale de distributions multi-modales et symétriques centrées sur le milieu de l'échelle ;

— il n'y a pas de différence flagrante entre les trois sessions ;

Influence du Groupe : Proximité de l'écran vs. « Sweet Spot »

La Fig. F.3 compare les distributions obtenues pour les deux groupes.

L'observation des histogrammes montre que nous observons à nouveau une tendance globale de distributions multi-modales symétriques centrées sur le milieu de l'échelle, sans différence flagrante entre les deux groupes.

Influence de la Balance : mixage A vs. mixage B

La Fig. F.4 compare les distributions obtenues pour les deux mixages A et B (le mixage

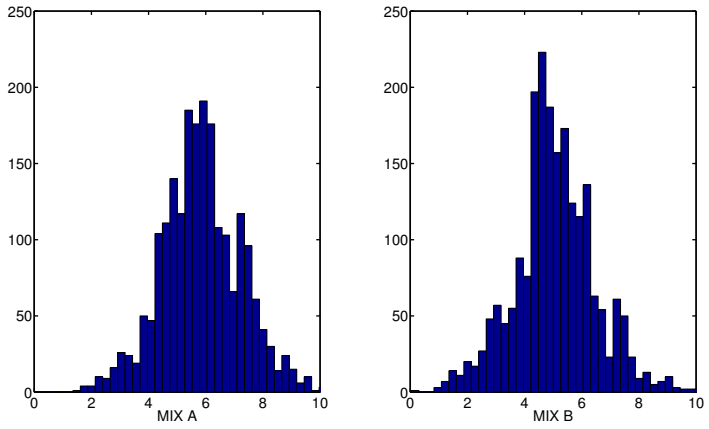


FIGURE F.4 – Comparaison des distributions obtenues pour le mixage A et pour le mixage B, toutes autres variables indépendantes confondues.

A ayant été mixé plus frontal par les expérimentateurs que le mixage B).

L'observation des histogrammes montre que :

— nous observons à nouveau une tendance globale de distributions multi-modales symétriques centrées sur le milieu de l'échelle ;

— les mixages A semblent avoir été notés légèrement plus frontaux que les mixages B, ce qui est logique.

Influence de la Séquence

La Fig. F.5 compare les distributions obtenues pour chacune des 8 séquences.

L'observation des histogrammes montre que :

— toutes les séquences présentent également une distribution symétrique centrée sur le milieu de l'échelle ;

— nous n'observons dans aucune séquence de distribution bi-modale marquée, qui aurait pu suggérer un effet prononcé de la stéréoscopie.

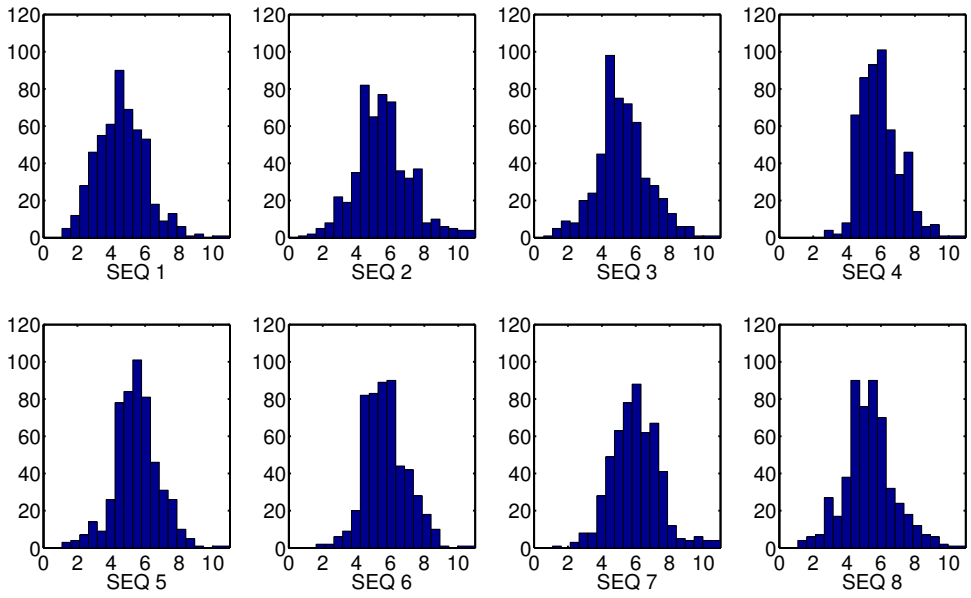


FIGURE F.5 – Comparaison des distributions obtenues pour chacune des 8 séquences, toutes autres variables indépendantes confondues.

Annexe G

Exploration des résultats de l'expérience III

Influence de la séquence

La Fig. G.1 compare les distributions obtenues pour les 8 séquences, toutes autres variables indépendantes confondues.

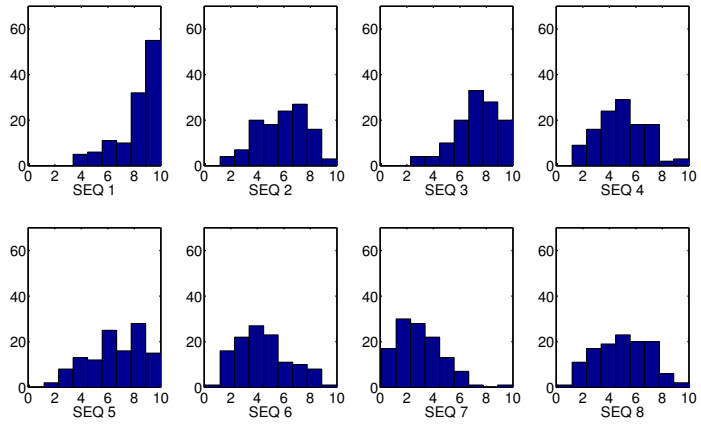


FIGURE G.1 – Comparaison des distributions obtenues pour chaque séquence dans l'expérience III, toutes autres variables indépendantes confondues.

L'observation des histogrammes montre bien que les différences perçues entre versions 2D et 3D-s ont été vraisemblablement différentes d'une séquence à l'autre. Les séquences 1 et 3, par exemple, semblent avoir donné lieu à des différences importantes, alors que les différences entre versions 2D et 3D-s paraissent beaucoup plus faibles dans la séquence 7. Certaines distributions semblent multi-modales, comme la séquence 5. Ces modes peuvent correspondre à différents groupes de sujets, ou alors à une influence de la session.

La Fig. G.2 compare les distributions obtenues pour chacun des deux groupes, toutes autres variables indépendantes confondues. La distribution pour le groupe 2 est symétrique et centrée sur le milieu de l'échelle, alors que celle pour le groupe 1 est plus plate.

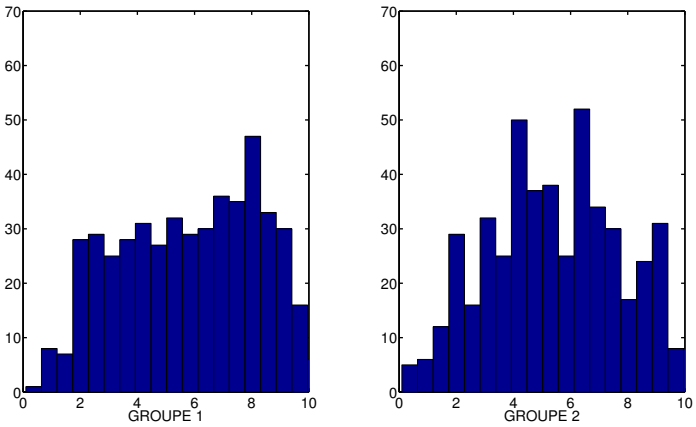


FIGURE G.2 – Comparaison des distributions obtenues pour chaque groupe dans l'expérience III, toutes autres variables indépendantes confondues.

La Fig. G.3 compare les distributions obtenues pour chacune des deux sessions, toutes autres variables indépendantes confondues. La distribution de la session 2 est symétrique et centrée sur le milieu de l'échelle. La session 1 semble multi-modale.

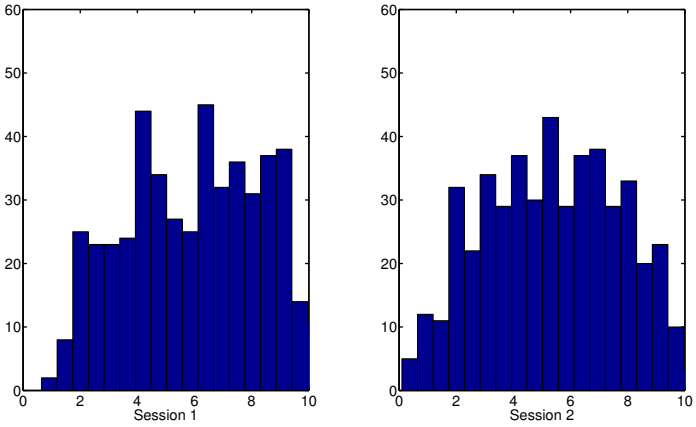


FIGURE G.3 – Comparaison des distributions obtenues pour chaque session dans l'expérience III, toutes autres variables indépendantes confondues.

Annexe H

Récapitulatif détaillé des séquences de l'expérience V

Séquence 1

Durée : 17 secondes

Plan fixe. Intérieur.

Dans cette séquence, une jeune femme est assise au centre d'un atelier sombre. Au fond, sur la droite, un homme arrive, ayant au téléphone une discussion animée avec son garagiste. L'homme semble plutôt agacé car le garagiste n'a toujours pas terminé la réparation de sa voiture. Il allume la lumière puis s'approche de la jeune fille. La jeune fille se lève et laisse sa place à l'homme, qui continue sa conversation téléphonique. L'homme s'assied et termine sa conversation.

Nous n'avons défini pour cette séquence qu'un seul objet : l'objet « homme », qui contient sa voix et ses bruits de pas. Nous avons choisi cette séquence car l'objet se déplace à la fois dans la profondeur (du fond du magasin au premier plan) et dans la latéralité (du bord-cadre droit de l'écran au centre de l'écran).

Les ambiances sont constituées de bruits de ventilation et de machines.

Pour le mixage « distance simulée », l'acoustique de la pièce a été simulée par une réverbération type « Medium Room ». La zone de brillance (6-7 kHz) ainsi que les fréquences graves et bas-mediums de l'objet « Homme » ont également été filtrées afin d'atténuer la sensation de présence de la prise de son de proximité.

Les tableaux ci-dessous présentent respectivement les niveaux sonores des pistes et les caractéristiques spatiales de l'objet pour le mixage « cohérent » (bien évidemment, azimuts et vitesses des objets sonores sont toujours nuls dans le cas des mixages « classiques » sur enceinte centrale). Un autre tableau sur la page suivante présente des captures d'écran illustrant le déroulement de la scène.

Piste audio	Niveau sonore moyen (dB RMS)
Objet « Homme » dans le mixage « proximité »	- 29.9
Objet « Homme » dans le mixage « distance simulée »	- 34.9
Ambiance L	- 53.7
Ambiance R	- 54.2
Ambiance Ls	- 55.2
Ambiance Rs	- 55.2

Tableau H.1 – Niveau sonore moyen de chaque piste audio pour la séquence 1

	Azimut moyen (valeur absolue)	Azimut max	Vitesse moyenne	Vitesse max
Objet « Homme »	12.5°	23°	3.7°/s	16.8°/s

Tableau H.2 – Caractéristiques spatiales de l'objet pour la séquence 1

SEQUENCE 1	
Une jeune femme est assise au centre d'un atelier.	
Au fond, sur la droite, un homme arrive, ayant au téléphone une discussion animée avec son garagiste. L'homme semble plutôt agacé car le garagiste n'a toujours pas terminé la réparation de sa voiture.	
Il allume la lumière.	
Il s'approche de la jeune fille.	
La jeune fille se lève et laisse sa place à l'homme, qui continue sa conversation téléphonique.	
L'homme s'assied et termine sa conversation.	

Tableau H.3 – Séquence 1

Séquence 2

Durée : 20 secondes

Travelling avant rapide, puis changement de plan sur un plan fixe. Extérieur.

Dans cette séquence, un cycliste est suivi de près par un conducteur pressé. La voiture finit par prendre une autre route. Le cycliste continue sa route et aboutit à une plage.

Il y a dans cette séquence deux objets sonores : le vélo (bruits de chaîne de vélo et du contact des pneus avec le gravier) et la voiture (bruits du moteur, crissements de pneu, bruits du contact des pneus avec le gravier). Nous avons choisi cette séquence car elle est dynamique, avec des déplacements rapides dans la profondeur (avec, par exemple, un changement de plan où le vélo se retrouve instantanément distant) et dans la latéralité (virage violent de la voiture du centre de l'écran vers la droite, traversée droite-gauche de l'écran par le vélo à la fin de la séquence).

Les ambiances sont constituées de bruits de vent, avec des vagues au loin et quelques oiseaux.

Pour le mixage « distance simulée », nous avons « creusé » dans les fréquences graves et bas-mediums, ainsi que dans la zone de brillance (6-7 kHz) afin d'atténuer la sensation de présence de la prise de son de proximité. Nous n'avons pas appliqué de réverbération sur cette séquence.

Les tableaux ci-dessous et sur la page suivante présentent respectivement les niveaux sonores des pistes, les caractéristiques spatiales des objets, ainsi que des captures d'écran illustrant le déroulement de la scène.

Piste audio	Niveau sonore moyen (dB RMS)
Objet « Vélo » dans le mixage « proximité »	-39.5
Objet « Vélo » dans le mixage « distance simulée »	-44.9
Objet « Voiture » dans le mixage « proximité »	-27.7
Objet « Voiture » dans le mixage « distance simulée »	-34.5
Ambiance L	-50.1
Ambiance R	-50.2
Ambiance Ls	- 57.1
Ambiance Rs	- 58.2

Tableau H.4 – Niveau sonore moyen de chaque piste audio pour la séquence 2

	Azimut moyen (valeur absolue)	Azimut max	Vitesse moyenne	Vitesse max
Objet « Vélo »	9.8°	23°	6.0°/s	58°/s
Objet « Voiture »	11.6°	23°	4.9°/s	22.7°/s

Tableau H.5 – Caractéristiques spatiales des objets pour la séquence 2. La valeur élevée de la vitesse maximale pour l'objet « Vélo » est due à un changement de plan.

SEQUENCE 2	
Un cycliste est suivi de près par un conducteur pressé.	
La voiture tourne a droite. Le cycliste l'insulte.	
Le cycliste continue sa route.	
Le cycliste se dirige vers une plage.	
Changement de plan : vue générale sur la plage. Le cycliste apparaît sur la droite et traverse de droite à gauche la plage.	

TABLEAU H.6 – Séquence 2

Séquence 3

Durée : 15 secondes

Plan fixe. Extérieur.

Dans cette séquence, un cycliste chute. Un homme tirant la barque apparaît et lui demande si tout va bien. Les deux hommes discutent. L'homme reprend sa barque tout en discutant puis s'éloigne. Le cycliste se retrouve seul sur la plage.

Il y a dans cette séquence trois objets sonores : la voix de l'« Homme 1 » (le cycliste), la voix de l'« Homme 2 » (l'homme à la barque) et la barque (bruits des chocs de la barque contre le sol). Nous avons choisi cette séquence car il s'agit d'une scène de dialogue avec des personnages au loin, en extérieur. Le déplacement droite-gauche de l'Homme 2 et de la barque est également intéressant.

Les ambiances sont constituées de bruits de vent, avec des vagues au loin et quelques oiseaux.

Pour le mixage « distance simulée », nous avons « creusé » dans les fréquences graves et bas-mediums, ainsi que dans la zone de brillance (6-7 kHz) afin d'atténuer la sensation de présence de la prise de son de proximité. Nous avons également ajouté une légère réverbération pour simuler les réflexions sur le sol et sur les rochers environnants.

Les tableaux ci-dessous et sur la page suivante présentent respectivement les niveaux sonores des pistes, les caractéristiques spatiales des objets, ainsi que des captures d'écran illustrant le déroulement de la scène.

Piste audio	Niveau sonore moyen (dB RMS)
Objet « Homme 1 » dans le mixage « proximité »	-37.2
Objet « Homme 1 » dans le mixage « distance simulée »	-41.7
Objet « Homme 2 » dans le mixage « proximité »	-29.6
Objet « Homme 2 » dans le mixage « distance simulée »	-33.3
Objet « Barque » dans le mixage « proximité »	-32.4
Objet « Barque » dans le mixage « distance simulée »	-41.8
Ambiance L	-49.2
Ambiance R	-51.8
Ambiance Ls	-58.3
Ambiance Rs	-61.2

Tableau H.7 – Niveau sonore moyen de chaque piste audio pour la séquence 3

	Azimut moyen (valeur absolue)	Azimut max	Vitesse moyenne	Vitesse max
Objet « Homme 1 »	0.6°	1.4°	1.0°/s	3.7°/s
Objet « Homme 2 »	7.6°	10°	4.3°/s	7.8°/s
Objet « Barque »	16.4°	23°	4.4°/s	10°/s

Tableau H.8 – Caractéristiques spatiales des objets pour la séquence 3

SEQUENCE 3	
Le cycliste chute. L'homme tirant la barque apparaît sur la droite.	
Les deux hommes discutent.	
L'homme reprend sa barque tout en discutant	
L'homme se dirige vers la gauche en tirant sa barque et en continuant de discuter	
Le cycliste se retrouve seul sur la plage.	

TABLEAU H.9 – Séquence 3

Séquence 4

Durée : 22 secondes

Plan fixe. Extérieur.

Dans cette séquence, une femme qui vient d'avoir un accident de voiture discute avec un enfant, au bord d'un ruisseau.

Il y a dans cette séquence deux objets sonores : la voix de la femme et la voix de l'enfant. Nous avons choisi cette séquence à cause de la différence de profondeur entre les deux objets (femme en gros plan, enfant légèrement en retrait) et d'azimut (femme sur la gauche de l'écran, enfant sur la droite).

Les ambiances sont constituées d'un fond d'air, avec un peu de vent, et d'une rivière très proche (on peut l'apercevoir en haut de l'écran sur la capture d'écran).

Pour le mixage « distance simulée », nous avons « creusé » dans les fréquences graves et bas-mediums, ainsi que dans la zone de brillance (6-7 kHz) afin d'atténuer la sensation de présence de la prise de son de proximité. Ce filtrage a été léger pour la femme, car celle-ci est très proche à l'écran, et bien plus prononcé pour l'enfant, car celui-ci est en retrait. Nous n'avons pas ajouté de réverbération.

Les tableaux ci-dessous et sur la page suivante présentent respectivement les niveaux sonores des pistes, les caractéristiques spatiales des objets, ainsi que des captures d'écran illustrant le déroulement de la scène.

Piste audio	Niveau sonore moyen (dB RMS)
Objet « Voix de la femme » dans le mixage « proximité »	-32.4
Objet « Voix de la femme » dans le mixage « distance simulée »	-33.2
Objet « Voix de l'enfant » dans le mixage « proximité »	-36.2
Objet « Voix de l'enfant » dans le mixage « distance simulée »	-42.3
Ambiance L	-54.0
Ambiance R	-53.8
Ambiance Ls	-60.8
Ambiance Rs	-60.3

Tableau H.10 – Niveau sonore moyen de chaque piste audio pour la séquence 4

	Azimut moyen (valeur absolue)	Azimut max	Vitesse moyenne	Vitesse max
Objet « Voix de la femme »	5.1°	7.2°	1.4°/s	4.9°/s
Objet « Voix de l'enfant »	12.1°	15.2°	1.1°/s	3.7°/s

Tableau H.11 – Caractéristiques spatiales des objets pour la séquence 4

FIGURE H.1 – Séquence 4

Séquence 5

Durée : 24 secondes

Plan fixe puis panoramique droite-gauche puis à nouveau plan fixe. Extérieur.

Dans cette séquence, un homme se balade dans une forêt. Il aperçoit un enfant jouant avec des branches. L'homme s'approche de l'enfant et discute avec lui. A deux reprises, l'enfant jette des branches sur l'homme en riant.

Il y a dans cette séquence trois objets sonores : la voix de l'homme, la voix de l'enfant et les branches (d'abord frappées l'une contre l'autre, puis jetées sur l'homme, avec des bruits d'impacts sur la tête de l'homme puis sur le sol). Nous avons choisi cette séquence à cause de la disposition gauche-droite de l'enfant et de l'homme lors du dialogue, ainsi que pour les mouvements rapides des branches , qui traversent l'écran de gauche à droite à deux reprises.

Les ambiances sont constituées de bruits de vents à travers des arbres.

Pour le mixage « distance simulée », nous avons « creusé » dans les fréquences graves et bas-mediums, ainsi que dans la zone de brillance (6-7 kHz) afin d'atténuer la sensation de présence de la prise de son de proximité. Nous avons également ajouté une légère réverbération pour simuler des réflexions sur les arbres et le sol.

Les tableaux ci-dessous et sur la page suivante présentent respectivement les niveaux sonores des pistes, les caractéristiques spatiales des objets, ainsi que des captures d'écran illustrant le déroulement de la scène.

Piste audio	Niveau sonore moyen (dB RMS)
Objet « Voix de l'homme » dans le mixage « proximité »	-42.9
Objet « Voix de l'homme » dans le mixage « distance simulée »	-45.5
Objet « Voix de l'enfant » dans le mixage « proximité »	-43.8
Objet « Voix de l'enfant » dans le mixage « distance simulée »	-44.8
Objet « Branches » dans le mixage « proximité »	-43.2
Objet « Branches » dans le mixage « distance simulée »	-45.9
Ambiance L	-67.7
Ambiance R	-68.1
Ambiance Ls	-75.0
Ambiance Rs	-75.1

Tableau H.12 – Niveau sonore moyen de chaque piste audio pour la séquence 5

	Azimut moyen (valeur absolue)	Azimut max	Vitesse moyenne	Vitesse max
Objet « Voix de l'homme »	4.4°	7.3°	1.3°/s	6.3°/s
Objet « Voix de l'enfant »	4.3°	5.6°	0.7°/s	1.6°/s
Objet « Branches »	20.1°	23°	1.4°/s	29.3°/s

Tableau H.13 – Caractéristiques spatiales des objets pour la séquence 5

SEQUENCE 5	
Un homme se balade dans une forêt. Il aperçoit un enfant jouant avec des branches (on peut apercevoir les branches dans le coin en bas à gauche de l'écran).	
L'homme s'approche de l'enfant.	
La caméra effectue un panoramique droite-gauche pour aboutir sur l'arbre dans lequel se trouve l'enfant.	
L'homme et l'enfant discutent. L'enfant jette des branches sur l'homme à deux reprises.	
L'enfant rigole.	

TABLEAU H.14 – Séquence 5

Séquence 6

Durée : 20 secondes

Plan fixe. Extérieur.

Dans cette séquence, un homme pousse un chariot dans un couloir en extérieur.

Il y a dans cette séquence un seul objet sonore : le chariot. Nous avons choisi cette séquence à cause du mouvement intéressant du chariot, aussi bien dans la latéralité que dans la profondeur : non seulement le chariot traverse l'écran de gauche à droite, mais en plus il part du premier plan pour ensuite s'éloigner progressivement.

Les ambiances sont constituées d'un fond d'air capté sur place.

Il n'y a pas de traitement ajouté pour cette séquence. Le mixage « distance simulée » a été obtenu à partir d'une prise de son en champ diffus. Le microphone était placé au niveau du bord-cadre gauche du plan.

Les tableaux ci-dessous et sur la page suivante présentent respectivement les niveaux sonores des pistes, les caractéristiques spatiales de l'objet, ainsi que des captures d'écran illustrant le déroulement de la scène.

Piste audio	Niveau sonore moyen (dB RMS)
Objet « Chariot » dans le mixage « proximité »	-40.1
Objet « Chariot » dans le mixage « distance simulée »	-45.0
Ambiance L	-48.0
Ambiance R	-48.1
Ambiance Ls	-57.3
Ambiance Rs	-57.4

Tableau H.15 – Niveau sonore moyen de chaque piste audio pour la séquence 6

	Azimut moyen (valeur absolue)	Azimut max	Vitesse moyenne	Vitesse max
Objet « Chariot »	14.9°	23°	1.1°/s	4.8°/s

Tableau H.16 – Caractéristiques spatiales de l'objet pour la séquence 6

SEQUENCE 6



Tableau H.17 – Séquence 6

Séquence 7

Durée : 16 secondes

Plan fixe. Intérieur.

Dans cette séquence, une femme fait la vaisselle en écoutant la radio.

Il y a dans cette séquence deux objets sonores : la vaisselle (chocs des ustensiles en train d'être lavés et reposés dans le lavabo) et la radio (diffusant de la publicité). Nous avons choisi cette séquence à cause de la latéralisation prononcée des deux objets. Cette séquence nous permet également d'avoir un exemple de situation avec deux objets à moyenne distance et en intérieur qui ne soient pas des dialogues.

Les ambiances sont constituées d'un fond d'air capté sur place.

Il n'y a pas de traitement ajouté pour cette séquence. Le mixage « distance simulée » a été obtenu à partir de prises de son en champ diffus. Pour la vaisselle, le microphone se situait à environ 3 m de la source. Pour la radio, nous avons légèrement rapproché le microphone car la couleur apportée par la salle nous avait paru peu naturelle à 3 m.

Les tableaux ci-dessous et sur la page suivante présentent respectivement les niveaux sonores des pistes, les caractéristiques spatiales des objets, ainsi que des captures d'écran illustrant le déroulement de la scène.

Piste audio	Niveau sonore moyen (dB RMS)
Objet « Vaisselle » dans le mixage « proximité »	-40.4
Objet « Vaisselle » dans le mixage « distance simulée »	-43.7
Objet « Radio » dans le mixage « proximité »	-40.2
Objet « Radio » dans le mixage « distance simulée »	-40.2
Ambiance L	-60.0
Ambiance R	-57.5
Ambiance Ls	-67.9
Ambiance Rs	-67.2

Tableau H.18 – Niveau sonore moyen de chaque piste audio pour la séquence 7

	Azimut moyen (valeur absolue)	Azimut max	Vitesse moyenne	Vitesse max
Objet « Vaisselle »	10°	10°	0°/s (statique)	0°/s (statique)
Objet « Radio »	18°	18°	0°/s (statique)	0°/s (statique)

Tableau H.19 – Caractéristiques spatiales des objets pour la séquence 7

FIGURE H.2 – Séquence 7

Séquence 8

Durée : 21 secondes

Plan fixe. Intérieur.

Dans cette séquence, une femme donne des instructions à un homme pour accrocher un tableau. Une fois la position déterminée, l'homme plante un clou dans le mur.

Il y a dans cette séquence deux objets sonores : la voix de la femme et le marteau (une quinzaine d'impacts du marteau sur le clou). Nous avons choisi cette séquence à cause de la latéralisation des objets (femme à gauche, marteau à droite) et de la différence de profondeur entre les objets (femme au premier plan, marteau en retrait).

Les ambiances sont constituées d'un fond d'air capté sur place.

Il n'y a pas de traitement ajouté pour cette séquence. Le mixage « distance simulée » a été obtenu à partir de prises de son en champ diffus. Pour la voix de femme, le microphone se situait à environ 2 m de la source. Pour le marteau, le microphone se situait à environ 3 m de la source.

Les tableaux ci-dessous et sur la page suivante présentent respectivement les niveaux sonores des pistes, les caractéristiques spatiales des objets, ainsi que des captures d'écran illustrant le déroulement de la scène.

Piste audio	Niveau sonore moyen (dB RMS)
Objet « Voix de la femme » dans le mixage « proximité »	-33.3
Objet « Voix de la femme » dans le mixage « distance simulée »	-32.6
Objet « Marteau » dans le mixage « proximité »	-31.5
Objet « Marteau » dans le mixage « distance simulée »	-36.0
Ambiance L	-43.7
Ambiance R	-47.5
Ambiance Ls	-62.2
Ambiance Rs	-60.2

Tableau H.20 – Niveau sonore moyen de chaque piste audio pour la séquence 8

	Azimut moyen (valeur absolue)	Azimut max	Vitesse moyenne	Vitesse max
Objet « Voix de la femme »	7.1°	11.0°	1.3°/s	3.2°/s
Objet « Marteau »	11.0°	11.0°	0°/s (statique)	0°/s (statique)

Tableau H.21 – Caractéristiques spatiales des objets pour la séquence 8

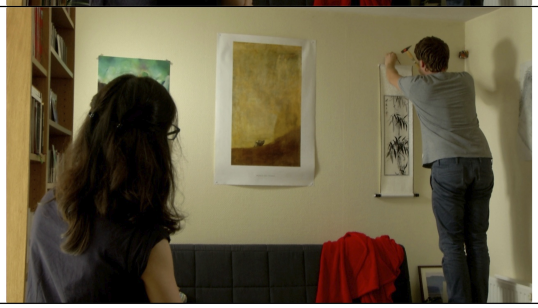
SEQUENCE 8	
Une femme donne des instructions à un homme pour correctement placer un tableau au mur.	
Une fois la position du tableau déterminée, l'homme plante un clou dans le mur avec un marteau.	

TABLEAU H.22 – Séquence 8

Annexe I

Exploration des résultats de l'expérience V

Effet du Mode Visuel : 3D-s vs. 2D

La Fig. I.1 compare les distributions obtenues pour les stimuli 3D-s et 2D, toutes autres variables indépendantes confondues.

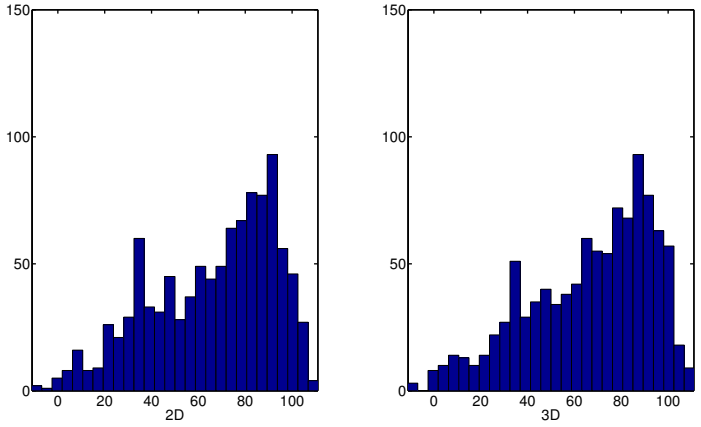


FIGURE I.1 – Comparaison des distributions obtenues pour le Mode Visuel 3D-s et 2D, toutes autres variables indépendantes confondues.

L'observation des histogrammes montre que :

— les notes sont resserrées vers le haut de l'échelle, notamment entre 80 et 100, ce qui suggère que les sujets ont globalement trouvé les bandes-son adaptées à l'image ;

— il n'y a pas à première vue de différence flagrante entre 2D et 3D-s. La stéréoscopie ne semble donc pas avoir influencé les jugements des sujets sur l'adéquation du son à l'image ;

— aussi bien en 2D qu'en 3D, un mode émerge dans la première moitié de l'échelle de notations (vers 30) ;

Effet de la Cohérence Azimutale : « classique » vs. « cohérent »

La Fig. I.2 compare les distributions obtenues pour les mixages « classiques » et « cohérents ».

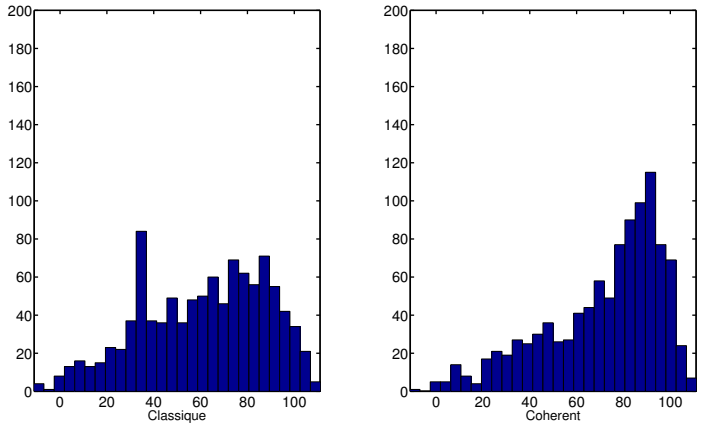


FIGURE I.2 – Comparaison des distributions obtenues pour les mixages « classiques » et « cohérents », toutes autres variables indépendantes confondues.

L'observation des histogrammes montre que :

— on retrouve le mode à 30 de la Fig. I.1 dans la distribution des mixages « classiques », mais pas dans la distribution des mixages « cohérents » ;

— L'allure globale des deux distributions montre des notes resserrées vers le haut de l'échelle, avec cependant un tassement vers le haut plus marqué pour les mixages « cohérents ». Les sujets semblent donc avoir préféré les mixages « cohérents » aux mixages « classiques ».

Effet de la Simulation de la Profondeur : « proximité » vs. « distance simulée »

La Fig. I.3 compare les distributions obtenues pour les mixages « proximité » et « distance simulée ».

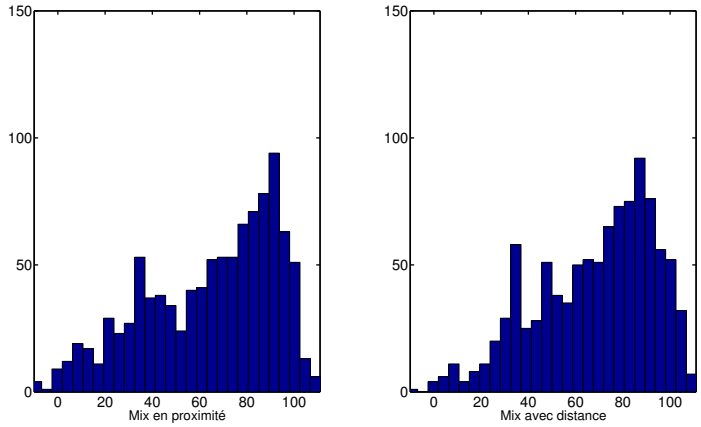


FIGURE I.3 – Comparaison des distributions obtenues pour les mixages « proximité » et « distance simulée », toutes autres variables indépendantes confondues.

L'observation des histogrammes montre que :

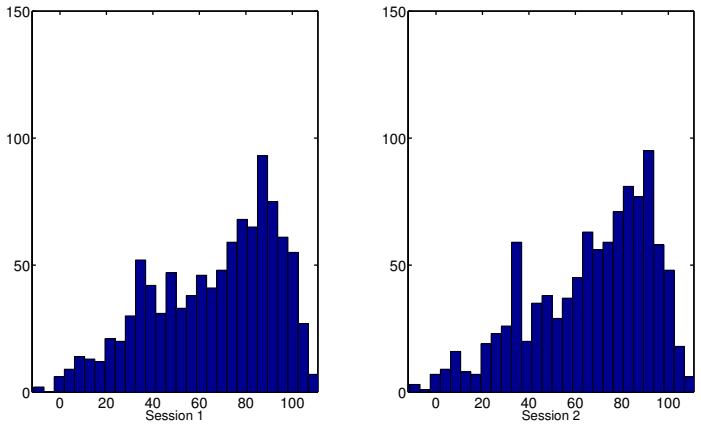


FIGURE I.4 – Comparaison des distributions obtenues pour les deux sessions, toutes autres variables indépendantes confondues.

— il n'y a pas de différence flagrante entre mixages « proximité » et mixages « distance simulée » ;
— on retrouve dans les deux distributions l'allure générale de tassement vers le haut de l'échelle. Ainsi, que la distance soit simulée ou non, les sujets semblent globalement trouver les bandes-son adaptées à l'image.
— on retrouve le mode à 30 dans les deux distributions.

Effet de la Répétition

La Fig. I.4 compare les distributions obtenues pour les deux sessions du test.

L'observation des histogrammes montre que :

— il n'y a pas de différence flagrante entre les deux sessions ;
— le mode à 30 est indépendant de la session.

Effet de la Cohérence Azimutale en fonction de la séquence

Une inspection des données suggère que le seul facteur simple pour lequel il y a un effet notable est la cohérence azimutale. La Fig. I.5 montre l'effet de la cohérence azimutale séquence par séquence.

L'observation des histogrammes montre que :

— la tendance globale de tassement vers le haut de l'échelle, déjà observée dans les distributions précédentes, se retrouvent dans les mixages « cohérents » pour toutes les séquences sauf pour la séquence 6 qui fait apparaitre deux modes. Une analyse plus en détail de la séquence 6 montre que cette allure semble être due au mixage cohérent en azimut « proximité » qui a été bien moins noté que le mixage cohérent en azimut « distance simulée » (voir Fig. I.6).

Nous avions observé sur la Fig. I.3 qu'il n'y avait pas de différence globale flagrante entre mixages « proximité » et mixages « distance simulée ». Cependant, le cas de la séquence 6 suggère qu'une préférence pour le mixage « distance simulée » par rapport

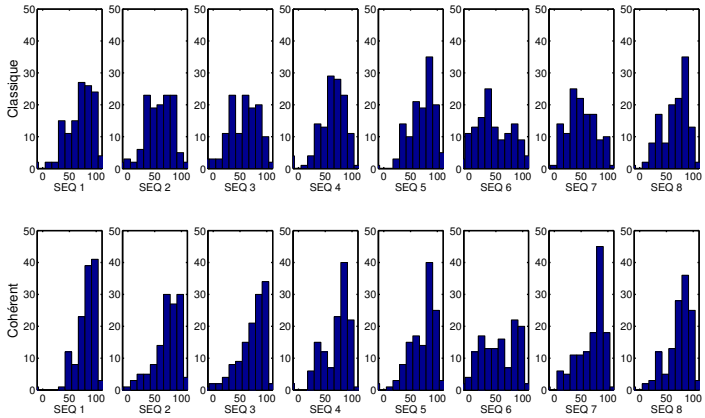


FIGURE I.5 – Distributions obtenues pour les mixages « classiques » et « cohérents », séquence par séquence.

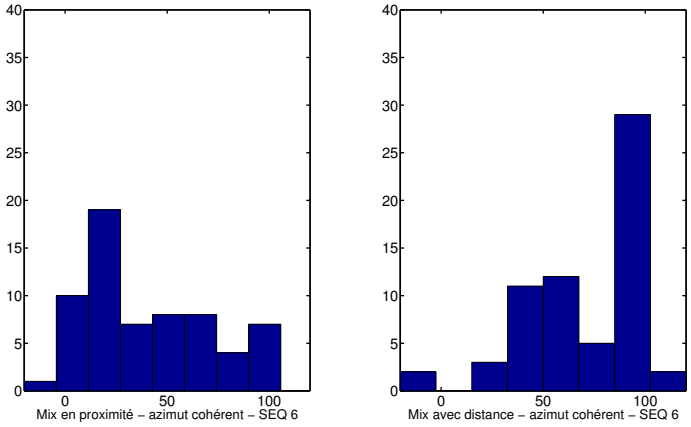


FIGURE I.6 – Distributions obtenues pour les mixages cohérents « proximité » et cohérent « distance simulée » de la séquence 6.

au mixage « proximité » est possible dans certains cas.

— Concernant les mixages « classiques », nous n'observons pas sur la Fig. I.5 de tassement vers le haut de l'échelle pour toutes les séquences, ce qui suggère que l'effet de la cohérence azimutale ne va pas être le même en fonction de la séquence.

— Le mode vers 30 n'est pas dû à une séquence en particulier qui aurait été particulièrement mal noté, puisqu'il émerge dans plusieurs séquences. Il semble plus probable que le mode soit dû à un sujet. Une analyse sujet par sujet montre que le mode est en fait dû au sujet n°8, qui a systématiquement noté dans le bas de l'échelle les mixages non cohérents. La figure I.7 montre la distribution obtenue pour les mixages « classiques » en enlevant les résultats obtenus par le sujet n°8 : on observe bien que le mode à 30 a presque disparu.

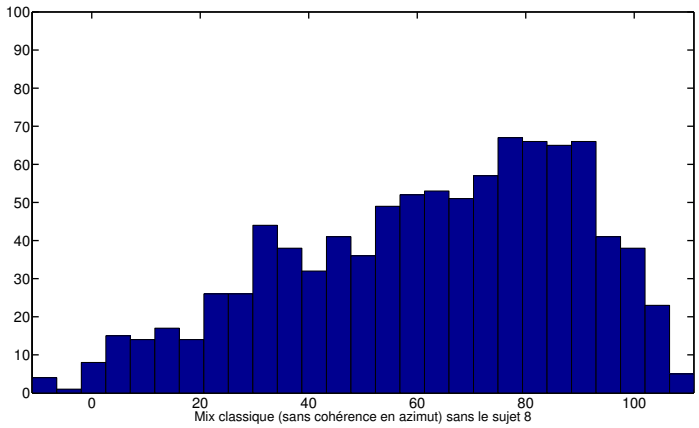


FIGURE I.7 – Distributions obtenues pour les mixages « classiques » sans le sujet n°8.

Bibliographie

Agganis, B. T., Muday, J. A. et Schirillo, J. A. (2010). Visual biasing of auditory localization in azimuth and depth. *Percept. Mot. Skills*, 111:872–892.

Alais, D. et Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Curr. Biol.*, 14:257–262.

Allen, I. (1991). Matching the sound to the picture. *In Proceedings of the 9th Audio Eng. Soc. Int. Conf. : Television Sound Today and Tomorrow.*

André, C. (2013). Audiovisual spatial congruence, and applications to 3d sound and stereoscopic video. *PhD thesis, Université de Liège, Belgique.*

André, C., Corteel, E., Embrechts, J.-J., Verly, J. et Katz, B. F. G. (2014). Subjective evaluation of the audiovisual spatial congruence in the case of stereoscopic-3d video and wave field synthesis. *Int. J. Hum.-Comput. St.*, 72:23–32.

André, C. R., Rébillat, M. et Katz, B. F. G. (2012). Sound for 3D cinema and the sense of presence. *In Proceedings of the 18th International Conference on Auditory Display.*

Ashmead, D. H., Leroy, D. et Odom, R. D. (1990). Perception of the relative distances of nearby sound sources. *Percept. Psychophys.*, 47:326–331.

Auro-3D (2006). Auro 3D 11.1 White Paper. http ://www.auro-3D.com/professional/technical-docs/.

Battaglia, P., Jacobs, R. et Aslin, R. (2003). Bayesian integration of visual and auditory signals for spatial localization. *J. Opt. Soc. Am.*, 20:1391–1396.

Bech, S. et Zacharov, N. (2006). *Perceptual Audio Evaluation.* John Wiley and Sons, Chichester.

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

Berkhout, A. J., de Vries, D. et Vogel, P. (1993). Acoustic control by Wave Field Synthesis. *J. Acoust. Soc. Am.*, 93:2764–2778.

Bermant, R. I. et Welch, R. B. (1976). Cross-modal bias and perceptual fusion with auditory-visual spatial discordance. *Percept. Mot. Skills*, 43:487–493.

Bertelson, P. et Aschersleben, G. (1998). Bayesian integration of visual and auditory signals for spatial localization. *Psychon. Bull. Rev.*, 5:482–489.

Bertelson, P. et Radeau, M. (1981). Cross-modal bias and perceptual fusion with auditory-visual spatial discordance. *Percept. Psychophys.*, 29:578–584.

Blauert, J. (1970). Ein versuch zum richtungshören bei gleichzeitiger optischer stimulation. *Acustica*, 23:118–119.

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

Blumlein, A. D. (1933). Improvements in and relating to sound-transmission, sound-recording and sound-reproducing systems. *British Patent*, 394:325.

Boone, M. M., Verheijen, E. N. G. et van Tol, P. F. (1995). Spatial sound-field reproduction by Wave-Field Synthesis. *J. Audio Eng. Soc.*, 43:1003–1012.

Bowen, A. L., Ramachandran, R., Muday, J. A. et Schirillo, J. A. (2011). Visual signals bias auditory targets in azimuth and depth. *Exp. Brain. Res.*, 214:403–414.

Bronkhorst, A. W. et Houtgast, T. (1999). Auditory distance perception in rooms. *Nature*, 397:517–520.

Bruijn, W. P. d. et Boone, M. M. (2002). Subjective experiments on the effects of combining spatialized audio and 2d video projection in audio-visual systems. *In Proceedings of the 112th Convention of the Audio Eng. Soc.* Paper no. 5582.

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

Brungart, D. S. et Rabinowitz, W. M. (1999). Auditory localization of nearby sources. head-related transfer functions. *J. Acoust. Soc. Am.*, 106:1465–1479.

Butler, R., Levy, E. T. et Neff, W. D. (1980). Apparent distance of sounds recorded in echoic and anechoic chambers. *J. Exp. Psychol.-Hum. Percept. Perform.*, 6:745–750.

Calcagno, E. R., Abregú, E. L., Eguia, M. C. et Vergara, R. (2012). The role of vision in auditory distance perception. *Percept.*, 41:175–192.

Carlile, S., Leong, P. et Hyams, S. (1997). The nature and distribution of errors in sound localization by human listeners. *Hearing Res.*, 114:179–196.

Cavonius, C. et Robbins, D. (1973). Relationships between luminance and visual acuity in the rhesus monkey. *J. Physiol.*, 232:239–246.

Chion, M. (2003). *Un art sonore, le cinéma.* Les Cahiers du Cinéma.

Chion, M. (2005). *L'audiovision.* Armand Colin.

Choe, C. S., Welch, R. B., Gilford, R. M. et Juola, J. F. (1975). The "ventriloquist effect" : Visual dominance or response bias ? *Percept. Psychophys.*, 18:55–60.

Coleman, P. D. (1968). Dual role of frequency spectrum in determination of auditory distance. *J. Acoust. Soc. Am.*, 44:631–632.

Coleman, M.. The sound of Avatar. http ://soundworkscollection.com/videos/avatar.

Coleman, M.. The sound of Hugo. http ://soundworkscollection.com/videos/hugo.

Corrigan, D., Gorzel, M., Squires, J. et Boland, F. (2013). Depth perception of audio sources in stereo 3d environments. *In Proc. SPIE 8648. Burlingame, CA.*

Corteel, E. et Nicol, R. (2003). Listening room compensation for wave field synthesis. what can be done ? *In Proceedings of the 23rd International Conference of the Audio Eng. Soc. : Signal Processing in Audio Recording and Reproduction.*

Coutant, B. E. et Westheimer, G. (1993). Population distribution of stereoscopic ability. *Ophthalmic Physiol. Opt.*, 13:3–7.

Czyzewski, A., Kornacki, A. et Odya, P. (2002). Some rules and methods for creation of surround sound. *In Proceedings of the 21st Conference of the Audio Eng. Soc.* Conference paper 000069.

Côté, N., Koehl, V. et Paquier, M. (2012). Ventriloquism effect on distance auditory cues. *In Proceedings of Acoustics 2012 joint congress (11ème Congrès Français d'Acoustique-2012 Annual IOA Meeting)*, pages 1063–1067.

Damaske, P. et Wagener, B. (1969). Richtungshörversuche über einen nachgebildeten kopf (investigations of directional hearing using a dummy head). *Acustica*, 21:30–35.

Djian, L. (2013). 3D : Eldorado ou échec ? *L'Express.* http ://www.lexpress.fr/culture/cinema/.

Dodgson, N. A. (2004). Variation and extrema of human interpupillary distance. *Proceedings of SPIE*, 5291:36–46.

Dolby (1994). Dolby Stereo Technical Guidelines for Dolby Stereo Theatres. http ://www.film-tech.com/warehouse/manuals/DOLBYTG1994.pdf.

Dolby (2011). Dolby Surround 7.1 Technical Paper. http ://www.dolby.com/us/en/technologies/dolby-surround-7-1-for-theater-tech-paper.pdf.

Dolby (2012). Next-Generation Audio for Cinema. http ://www.dolby.com/us/en/technologies/dolby-atmos/dolby-atmos-next-generation-audio-for-cinema-white-paper.pdf.

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

Dupas, A. (2007). multicanal au cinéma : un canal zénithal ? *Mémoire de fin d'études, École Nationale Supérieure Louis Lumière.*

Erkelens, C. J. et van Ee, R. (2002). The role of the cyclopean eye in vision : sometimes inappropriate, always irrelevant. *Vision Res.*, 42:1157–1163.

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

Gambier, V. (2010). Nouvelle approche pour la bande sonore d'un film en relief. *Mémoire de fin d'études, École Nationale Supérieure Louis Lumière*, page 55.

Gardner, M. (1969). Distance estimation of 0° or apparent 0°-oriented speech signals in anechoic space. *J. Acoust. Soc. Am.*, 45:47–53.

Gardner, M. B. (1968). Proximity image effect in sound localization. *J. Acoust. Soc. Am.*, 43:163.

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

Giard, M. H. et Peronet, F. (1999). Auditory-visual integration during multimodal object recognition in humans : A behavioral and electrophysiological study. *J. Cognitive Neurosci.*, 11:473–490.

Girden, E. R. (1991). *ANOVA : Repeated measures.* Sage Publications Inc, Newbury Park.

Green, S. B. et Salkind, N. J. (2013). *Using SPSS for Windows and Macintosh.* Pearson, septième édition.

Griesinger, D. (2002). Stereo and surround panning in practice. *In Proceedings of the 112th Audio Eng. Soc. Conv.*

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.*

Hairston, W., Wallace, M., Vaughan, J., Stein, B., Norris, J. et Schirillo, J. (2003). Visual localization ability influences cross-modal bias. *J. Cogn. Neurosci.*, 15:20–29.

Hamasaki, K., Hatano, W. et Hiyama, K. (2004). 5.1 and 22.2 multichannel sound productions using an integrated surround sound panning system. *In Proceedings of the 117th Audio Eng. Soc. Conv.*

Haustein, B. G. (1969). Hypothesen über die einohrige entfernungswahrnehmung des menschlichen gehörs (hypotheses about the perception of distance in human hearing with one ear). *Hochfrequenztech. u. Elektroakustik*, 78:46–57.

Haustein, B. G. et Schirmer, W. (1970). Messeinrichtung zur untersuchung des richtungslokalisationsvermögens (a measuring apparatus for the investigation of the faculty of directional localization). *Hochfrequenztech. u. Elektroakustik*, 79:96–101.

Hays, W. L. (1994). *Statistics.* Harcourt Brace, Fort Worth, cinquième édition.

Hendrickx, E., Paquier, M. et Koehl, V. (2014). The influence of stereoscopy on the sound mixing of movies : A study on the front/rear balance of ambience. *J. Audio Eng. Soc.*, 62:723–735.

Hendrickx, E., Paquier, M. et Koehl, V. (2015). Audiovisual spatial coherence for 2D and stereoscopic-3D movies. *J. Audio Eng. Soc.* (Accepté).

Hering, E. (1942). *Spatial sense and movements of the eye.* Baltimore, MD : American Academy of Optometry.

Hiekkanen, T., Lempiainen, T., Mattila, M., Veijanen, V. et Pulkki, V. (2007). Reproduction of virtual reality with multichannel microphone techniques. *In Proceedings of the 122nd Convention of the Audio Eng. Soc.* convention paper 7070.

Hládek, L., Le Dantec, C. C., Kopco, N. et Seitz, A. (2013). Ventriloquism effect and aftereffect in the distance dimension. *In Proceedings of Meetings on Acoustics, Acoust. Soc. Am.*, volume 19. p. 050042.

Hollander, M. et Wolfe, D. A. (1999). *Nonparametric Statistical Methods.* John Wiley and Sons, Hoboken, deuxième édition.

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

Howell, D. C. (2009). *Statistical Methods for Psychology.* Cengage Learning, Wadsworth, septième édition.

Hugonnet, C. et Jouhaneau, J. (1987). Comparative spatial transfer function of six different stereophonic systems. *In Proceedings of the 82nd Convention of the Audio Eng. Soc.* convention paper 2465.

Hugonnet, C. et Walder, P. (1995). *Théorie et pratique de la prise de son stéréophonique.* Eyrolles.

IEC 60268-13 (1998). Sound system equipment - Part 13 : Listening tests on loudspeakers. International Electrotechnical Commission.

Ijsselsteijn, W., de Ridder, H., Freeman, J., Avons, S. E. et Bouwhuis, D. (2001). Effects of stereoscopic presentation, image motion, and screen size on subjective and objective corroborative measures of presence. *Presence-Teleop. Virt.*, 10:298–311.

Iljazovic, A., Leschka, F., Neugebauer, B. et Plogsties, J. (2012). The influence of 2-D and 3-D video playback on the perceived quality of spatial audio rendering for headphones. *In Proceedings of the 133rd Convention of the Audio Eng. Soc.* convention paper 8735.

ITU-R BS.1116-1 (1994). Methods for Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems. International Telecommunications Union.

ITU-R BS.1284-1 (2003). General methods for the subjective assessment of sound quality. International Telecommunications Union.

ITU-R BS.1286 (1997). Methods for the subjective assessment of audio systems with accompanying picture. International Telecommunications Union.

Jack, C. E. et Thurlow, W. R. (1973). Effects of degree of visual association and angle of displacement on the "ventriloquism" effect. *Percept. Mot. Skills*, 37:967–979.

Jackson, C. V. (1953). Visual factors in auditory localization. *Q. J. Exp. Psychol.*, 5:52–65.

Jones, C. (2009). Roll-out and financial aspects of 3d digital cinema. *In Proceedings of 3D Stereo MEDIA 2009. Liège, Belgium.*

Jouhaneau, J. (2012). Perception de l'espace et immersion - perception visuelle. *Techniques de l'ingénieur.*

Jullier, L. (2006). *Le son au cinéma.* Les Cahiers du cinéma.

Kamekawa, T., Marui, A. et Enatsu, M. (2011). Evaluation of spatial impression comparing 2ch stereo, 5ch surround, and 7ch surround with height channels for 3d imagery. *In Proceedings of the 130th Audio Eng. Soc. Conv.*

Knudsen, E. I. et Konishi, M. (1979). Mechanisms of sound localization in the barn owl (tyto alba). *J. Comp. Physiol.*, 133:13–21.

Komiyama, S. (1989). Subjective evaluation of angular displacement between picture and sound directions for HDTV sound systems. *J. Audio Eng. Soc.*, 37:210–214.

Krohn, B. (2009). Entretien avec Paul Martin Smith. *Les Cahiers du Cinéma*, 647:16–19.

Kruszielski, L. F., Kamekawa, T. et Marui, A. (2012). Perception of distance and the effect on sound recording distance suitability for a 3D or 2D image. *In Proceedings of the 133rd Convention of the Audio Eng. Soc.* eBrief 55.

Kunka, B. et Kostek, B. (2013). New aspects of virtual sound source localization research-impact of visual angle and 3-D video content on sound perception. *J. Audio Eng. Soc.*, 61:280–289.

Lam, C. F., Dubno, J. R. et Mills, J. H. (1999). Determination of optimal data placement for psychometric function estimation : a computer simulation. *J. Acoust. Soc. Am.*, 106:1969–1976.

Letowski, T. et Letowski, S. (2011). Localization error : Accuracy and precision of auditory localization. *Advances in Sound Localization.*

Lewald, J., Ehrenstein, W. H. et Guski, R. (2001). Spatio-temporal constraints for auditory-visual integration. *Behav. Brain Res.*, 121:69–79.

Makous, J. C. et Middlebrooks, J. C. (1990). Two-dimensional sound localization by human listeners. *J. Acoust. Soc. Am.*, 87:2188–2200.

Mannerheim, P. (2011). Spatial sound and stereoscopic vision. *In Proceedings of the 130th Convention of the Audio Eng. Soc.* Paper no. 8424.

Martin, R. L., McAnally, K. I., Bolia, R. S., Eberle, G. et Brungart, D. S. (2012). Spatial release from speech-on-speech masking in the median sagittal plane. *J. Acoust. Soc. Am.*, 131:378–385.

Melchior, F., Brix, S., Sporer, T., Roder, T. et Klehs, B. (2003). Wave field synthesis in combination with 2D video projection. *In Proceedings of the 24th International Conference of the Audio Eng. Soc.* Paper no. 47.

Melchior, F., Fischer, J. et de Vries, D. (2006). Audiovisual perception using wave field synthesis in combination with augmented reality systems : Horizontal positioning. *In Proceedings of the 28th International Conference of the Audio Eng. Soc.* Paper no. 3-2.

Mendiburu, B. (2009). *3D Movie Making : Stereoscopic Digital Cinema from Script to Screen*. Focal Press.

Mershon, D. et King, L. E. (1975). Intensity and reverberation as factors in the auditory perception of egocentric distance. *Atten. Percept. Psychophys.*, 18:409–415.

Mershon, D. H. et Bowers, J. N. (1979). Absolute and relative cues for the auditory perception of egocentric distance. *Percept.*, 8:311–322.

Mershon, D. H., Desaulniers, D. H., Amerson, T. L. et Kiefer, S. A. (1980). Visual capture in auditory distance perception : Proximity image effect reconsidered. *J. Aud. Res.*, 20:129–136.

Middlebrooks, J. C. (1992). Narrow-band sound localization related exter- nal ears acoustics. *J. Acoust. Soc. Am.*, 92:2607–2624.

Middlebrooks, J. C. et Green, D. M. (1991). Sound localization by human listeners. *Annu. Rev. Psychol.*, 42:135–159.

Middlebrooks, J. C., Makous, J. et Green, D. M. (1989). Directional sensitivity of sound-pressure levels in the human ear canal. *J. Acoust. Soc. Am.*, 86:89–108.

Moore, B. (2012). *An introduction to the psychology of hearing*. Brill.

Moulin, S. (2015). Quel son spatialisé pour la vidéo 3D ? influence d'un rendu Wave Field Synthesis sur l'expérience audio-visuelle 3D. *PhD thesis, Université Paris Descartes*.

Moulin, S., Nicol, R., Gros, L. et Mamassian, P. (2013). nfluence of the audio rendering on 3D audiovisual experience. *Acoustics in Practice*, 1:29–36.

Nielsen, S. H. (1993). Auditory distance perception in different rooms. *J. Audio Eng. Soc.*, 41:755–770.

Nunnally, J. C. et Bernstein, I. H. (1994). *Psychometric Theory*. McGraw-Hill, troisième édition.

Oldfield, S. et Parker, S. (1984). Acuity of sound localization : a topography of auditory space. i. normal hearing conditions. *Percept.*, 13:581–600.

Perrott, D. et Saberi, K. (1990). Minimum audible angle thresholds for sources varying in both elevation and azimuth. *J. Acoust. Soc. Am.*, 87:1728–1731.

Pick, H. L., Warren, D. H. et Hay, J. C. (1969). Sensory conflict in judgments of spatial direction. *Percept. Psychophys.*, 6:203–205.

Poulton, E. C. (1992). *Bias in quantifying judgments.* Lawrence Erlbaum Associates, Hillsdale.

Preibisch-Effenberger, R. (1966). Die schallokalisationsfähigkeit des menschen und ihre audiometrische verwendbarkeit zur klinischen diagnostik (the human faculty of sound localization and its audiometric application to clinical diagnostics). *Habilitationsschrift, Medizinische Akademie, Dresden.*

Première (2013). Gravity ou le triomphe de la 3D. *Première.* http ://www.premiere.fr/Cinema/News-Cinema/Gravity-ou-le-triomphe-de-la-3D-3865273.

Pulkki, V. et Karjalainen, M. (2001). Localization of amplitude-panned virtual sources i : Stereophonic panning. *J. Audio Eng. Soc.*, 49:739–752.

Radeau, M. (1974). Adaptation au déplacement prismatique sur la base d'une discordance entre la vision et l'audition. *L'Année Psychologique*, 74:23–24.

Radeau, M. et Bertelson, P. (1976). The effect of a textured visual field on modality dominance in a ventriloquism situation. *Percept. Psychophys.*, 20:227–235.

Radeau, M. et Bertelson, P. (1977). Adaptation to auditory-visual discordance and ventriloquism in semirealistic situations. *Percept. Psychophys.*, 22:137–146.

Recanzone, G. H., Makhamra, S. D. et Guard, D. C. (1998). Comparison of relative and absolute sound localization ability in humans. *J. Acoust. Soc. Am.*, 103:1085–1097.

Recommandation RT 012 (2003). Salles de spectacle cinématographique Confort du Spectateur. Commission Supérieure Technique de l'Image et du Son, Paris.

Recommandation RT 013 (2006). Niveau sonore en salle. Commission Supérieure Technique de l'Image et du Son, Paris.

Rohr, L., Corteel, E., Nguyen, K. V. et Lissek, H. (2013). Vertical localization performance in a practical 3-D WFS formulation. *J. Audio Eng. Soc.*, 61:1001–1014.

Rumsey, F. (2002). Spatial quality evaluation for reproduced sound : Terminology, meaning, and a scene-based paradigm. *J. Audio Eng. Soc.*, 50:651–666.

Rébillat, M., Boutillon, X., Corteel, É. et Katz, B. (2012). Audio, visual, and audio-visual egocentric distance perception by moving subjects in virtual environments. *ACM Trans. Appl. Percept.*, 9:19 :1–19 :17.

Rébillat, M., Corteel, É. et Katz, B. (2008). Smart-I2 : Spatial multi-user audio-visual real time interactive interface. *In Proceedings of the 125th Audio Eng. Soc. Conv.*

Seeber, B. U. et Fastl, H. (2004). On auditory-visual interaction in real and virtual environments. *In Proc. ICA 2004*, pages 2293–2296.

Shinn-Cunningham, B. G. (2000). Distance cues for virtual auditory space. *In Proceedings of the IEEE Pacific-Rim Conference on Multimedia*, pages 227–230.

Shinn-Cunningham, B. G., Kopco, N. et Martin, T. J. (2005). Localizing nearby sound sources in a classroom : Binaural room impulse responses. *J. Acoust. Soc. Am.*, 117:3100–3115.

Spector, R. H. (1990). *Visual Fields - In : Clinical Methods : The History, Physical, and Laboratory Examinations.* Walker HK, Hall WD, Hurst JW, editors.

Sporer, T. (2004). Wave Field Synthesis-generation and reproduction of natural sound environments. *In Proceedings of the 7th International Conference on Digital Audio Effects (DAFx-04), Naples, Italy.*

Standard AES20-1996 (r2007) (1996, reaffirmed 2007, stabilized 2008). AES recommended practice for professional audio - subjective evaluation of loudspeakers. Audio Engineering Society.

Start, E. (1997). Direct sound enhancement by Wave Field Synthesis. *PhD thesis, Delft University of Technology.*

Stern, R. M., Brown, G. J. et Wang, D. (2006). *Binaural sound localization. In : Computational Auditory Scene Analysis : Principles, Algorithms, and Applications.* Wiley-IEEE Press.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, 64:153–181.

Stewart, G. W. (1920). The function of intensity and phase in the binaural location of pure tones. II. *Physical Review*, 15:432–445.

Streicher, R. et Everest, F. A. (2006). *The new stereo soundbook.* Audio Engineering Associates, Pasadena, troisième édition.

Theile, G. (Feb. 2000). Multichannel natural recording based on psychoacoustic principles. *In Proceedings of the 108th Convention of the Audio Eng. Soc.* convention paper 5156.

Thurlow, W. R. et Jack, C. E. (1973). Certain determinants of the "ventriloquism effect". *Percept. Mot. Skills*, 36:1171–1184.

Toole, F. E. (2008). *Sound Reproduction.* Focal Press, Burlington.

Turner, A., Berry, J. et Holliman, N. (2011). Can the perception of depth in stereoscopic images be influenced by 3D sound ? *In Proceedings of SPIE 7863, 786307.*

Vatakis, A. et Spence, C. (2007). Crossmodal binding : Evaluating the unity assumption using audiovisual speech stimuli. *Percept. Psychophys.*, 69:744–756.

Verheijen, E. (1998). Sound reproduction by Wave Field Synthesis. *PhD thesis, Delft University of Technology.*

Vogel, P. (1993). Application of Wave Field Synthesis in room acoustics. *PhD thesis, Delft University of Technology.*

Wallace, M., Roberson, G., Hairston, W., Stein, B., Vaughan, J. et Schirillo, J. (2004). Unifying multisensory signals across time and space. *Exp. Brain. Res.*, 158:252–258.

Warren, D. H. (1979). Spatial localization under conflict conditions : Is there a single explanation ? *Percept.*, 8:323–337.

Warren, D. H., Welch, R. B. et McCarthy, T. J. (1981). The role of visual-auditory compellingness in the ventriloquism effect : Implications for transitivity among the spatial senses. *Percept. Psychophys.*, 30:557–564.

Warren, R. M. (1999). *Auditory Perception. A New Analysis and Synthesis.* Cambridge University Press.

Weerts, T. C. et Thurlow, W. R. (1971). The effect of eye position and expectation on sound localization. *Percept. Psychophys.*, 9:35–39.

Welch, R. B. (1999). Meaning, attention, and the "unity assumption" in the intersensory bias of spatial and temporal perceptions. *Adv. Psychol.*, 129:371–387.

Welch, R. B. et Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychol. Bull.*, 88:638.

Werner, S., Liebetrau, J. et Sporer, T. (2013). Vertical sound source localization influenced by visual stimuli. *Signal Process. Res.*, 2:29–38.

Wettschurek, R. (1970). Über unterschiedsschwellen beim richtungshoren in der medianebene. *Gemeinschaftstagung für Akustik und Schwingungstechnik*, pages 385–388.

Wittek, H., Haut, C. et Keinath, D. (2006). Double M/S - a surround technique put to test. *Tonmeistertagung*, pages 1–3.

Woods, A. J. (2001). Optimal usage of LCD projectors for polarized stereoscopic projection. *Photonics West 2001-Electronic Imaging. International Society for Optics and Photonics*, pages 5–7.

Woodworth, R. S. et Schlosberg, H. (1962). *Experimental psychology.* Oxford and IBH Publishing.

Zahorik, P. (2001). Estimating sound source distance with and without vision. *Optom. Vis. Sci.*, 78:270–275.

Zahorik, P. (2002a). Assessing auditory distance perception using virtual acoustics. *J. Acoust. Soc. Am.*, 111:1832–1846.

Zahorik, P. (2002b). Auditory display of sound source distance. *In Proceedings of the 8th International Conference on Auditory Displays*, pages 326–332.

Zahorik, P. (2003). Auditory and visual distance perception : The proximity-image effect revisited. *J. Acoust. Soc. Am.*, 113:2270–2270.

Zahorik, P., Brungart, D. S. et Bronkhorst, A. W. (2005). Auditory distance perception in humans : A summary of past and present research. *Acta Acust. united Ac.*, 91:409–420.

Zchaluk, K. et Foster, D. H. (2009). Model-free estimation of the psychometric function. *Percept. Psychophys.*, 71:1414–1425.

Zielinski, S., Rumsey, F. et Bech, S. (2003). Effects of down-mix algorithms on quality of surround sound. *J. Audio Eng. Soc.*, 51:780–798.

Zielinski, S., Rumsey, F. et Bech, S. (2008). On some biases encountered in modern audio quality listening tests - a review. *J. Audio Eng. Soc.*, 56:427–451.

Résumé

Peu d'études ont été menées sur l'influence de la stéréoscopie sur la perception d'un mixage audio au cinéma. Les témoignages de mixeurs ou les articles scientifiques montrent pourtant une grande diversité d'opinions à ce sujet. Certains estiment que cette influence est négligeable, d'autres affirment qu'il faut totalement revoir notre conception de la bande-son, aussi bien au niveau du mixage que de la diffusion.
Une première série d'expériences s'est intéressée à la perception des sons d'ambiance. 8 séquences, dans leurs versions stéréoscopique (3D-s) et non-stéréoscopique (2D), ont été diffusées dans un cinéma à des sujets avec plusieurs mixages différents. Pour chaque présentation, les sujets devaient évaluer à quel point le mixage proposé leur paraissait trop frontal ou au contraire trop «*surround*», le but étant de mettre en évidence une éventuelle influence de la stéréoscopie sur la perception de la balance avant/arrière d'un mixage audio. Les résultats obtenus ont rejoint ceux d'une expérience préliminaire menée dans un auditorium de mixage, où les sujets se trouvaient en situation de mixeur et devaient eux-mêmes régler la balance avant/arrière : l'influence de la stéréoscopie était faible et n'apparaissait que pour quelques séquences. Une troisième expérience fut conduite pour vérifier si les séquences pour lesquelles la perception de la balance avant/arrière était significativement impactée par la stéréoscopie étaient celles dont les différences entre versions 2D et 3D-s étaient les plus importantes en termes de profondeur visuelle perçue. Cependant, aucune corrélation n'a pu être trouvée.
Des études ont ensuite été menées sur la perception des objets sonores tels que dialogues et effets. Une quatrième expérience s'est intéressée à l'effet ventriloque en élévation : lorsque l'on présente à un sujet des stimuli audio et visuel temporellement coïncidents mais spatialement disparates, les sujets perçoivent parfois le stimulus sonore au même endroit que le stimulus visuel. On appelle ce phénomène l'*effet ventriloque* car il rappelle l'illusion créée par le ventriloque lorsque sa voix semble plutôt provenir de sa marionnette que de sa propre bouche. Ce phénomène a été très largement étudié dans le plan horizontal, et dans une moindre mesure en distance. Par contre, très peu d'études se sont intéressées à l'élévation. Dans cette expérience, nous avons présenté à des sujets des séquences audiovisuelles montrant un homme en train de parler. Sa voix pouvait être reproduite sur différents haut-parleurs, qui créaient des disparités plus ou moins grandes en azimut et en élévation entre le son et l'image. Pour chaque présentation, les sujets devaient indiquer si la voix semblait ou non provenir de la même direction que la bouche de l'acteur. Les résultats ont montré que l'effet ventriloque était très efficace en élévation, ce qui suggère qu'il n'est peut-être pas nécessaire de rechercher la cohérence audiovisuelle en élévation au cinéma.
Une cinquième et dernière expérience a permis d'étudier l'influence de la stéréoscopie sur les attentes des spectateurs en termes de cohérence audiovisuelle spatiale. Au cinéma, les objets sonores sont en général diffusés sur l'enceinte centrale, indépendamment de la position à l'écran des sources visuelles associées. Cependant, certains ingénieurs du son et chercheurs ont suggéré que la cohérence audiovisuelle spatiale pouvait améliorer significativement l'expérience des spectateurs, surtout dans le cas de films en relief. Dans cette expérience, les sujets devaient évaluer à quel point la bande-son leur paraissait « adaptée » à l'image pour 8 séquences projetées en 2D et en 3D-s. Selon la bande-son, les sources sonores pouvaient être plus ou moins cohérentes en azimut et en profondeur avec la position de leur source visuelle respective sur l'écran (la cohérence en élévation avait été mise de côté au vu des résultats de l'expérience 4). Les résultats ont montré que la cohérence en azimut pouvait améliorer significativement l'adéquation du son à l'image. En profondeur, une amélioration a pu être constatée, mais seulement pour une séquence. Par contre, la stéréoscopie n'a eu aucune influence sur les jugements des sujets, en accord avec les résultats des premières expériences sur la perception des sons d'ambiance.

Abstract

Few psychoacoustic studies have been carried out about the influence of stereoscopy on the sound mixing of movies. Yet very different opinions can be found in the cinema industry and in scientific papers. Some argue that sound needs to be mixed differently for stereoscopic movies while others pretend that this influence is negligible.
A first set of experiments was conducted, which focused on the perception of ambience. Eight sequences - in their stereoscopic (s-3D) and non-stereoscopic (2D) versions, with several different sound mixes - were presented to subjects. For each presentation, subjects had to judge to what extent the mix sounded frontal or « surround. » The goal was to verify whether stereoscopy had an influence on the perception of the front/rear balance of ambience. Results showed that this influence was weak, which was consistent with a preliminary experiment conducted in a mixing auditorium where subjects had to mix the front/rear balance of several sequences themselves.
A third experiment was conducted to verify if the sequences that were significantly influenced by stereoscopy corresponded to the sequences whose differences between s-3D and 2D versions were the most important in terms of perceived visual depth, yet no correlation could be found.
Studies were then conducted on the perception of sound objects such as dialogs or on-screen effects. A fourth experiment focused on ventriloquism in elevation: when presented with a spatially discordant auditory-visual stimulus, subjects sometimes perceive the sound and the visual stimuli as coming from the same location. Such a phenomenon is often referred to as *ventriloquism*, because it evokes the illusion created by a ventriloquist when his voice seems to emanate from his puppet rather than from his mouth. While this effect has been extensively examined in the horizontal plane and to a lesser extent in distance, few psychoacoustic studies have focused on elevation. In this experiment, sequences of a man talking were presented to subjects. His voice could be reproduced on different loudspeakers, which created disparities in both azimuth and elevation between the sound and the visual stimuli. For each presentation, subjects had to indicate whether or not the voice seemed to emanate from the mouth of the actor. Ventriloquism was found to be highly effective in elevation, which suggests that audiovisual coherence in elevation might be unnecessary in theaters.
In a fifth experiment, the influence of stereoscopy on subjects' expectations regarding audiovisual spatial coherence was investigated. In theaters, sound objects are most of the time reproduced on the central loudspeaker, regardless of the position on screen of their related visual sources. Yet, some sound engineers and researchers have suggested that a spatial audiovisual coherence could improve the experience of the audience significantly, especially for s-3D movies. In this experiment, subjects were asked to judge the suitability of several soundtracks for 8 sequences, which were presented in their s-3D and 2D versions. Depending on the soundtrack, sound sources could be more or less coherent in azimuth and in depth to their related visual sources (coherence in elevation was not investigated because of the results of the fourth experiment). Results showed that sound suitability could be significantly improved for most of the sequences when coherence in azimuth was achieved. In depth, improvement was only observed with one sequence. However, no significant effect of stereoscopy on subjects' judgments could be found, which is consistent with the previous experiments on the perception of ambience.