ÉCOLE NATIONALE SUPÉRIEURE LOUIS LUMIÈRE

MÉMOIRE DE FIN D'ÉTUDES

Cohérence audiovisuelle spatiale au cinéma : Influence de la vitesse et de la nature des objets sonores

Clément TIJOU
Section son Promotion 2016

Directeur de mémoire interne :
Etienne HENDRICKX
Directeur de mémoire externe :
Simon APOSTOLOU
Rapporteur :
Sylvain LAMBINET

Date de soutenance : 13 Juin 2016

Remerciements

Avant de présenter mon travail je tiens à remercier tous ceux qui ont permis à ce mémoire de voir le jour.

Je tiens d'abord à remercier Etienne Hendrickx et Simon Apostolou qui m'ont accompagné tout au long de ce projet, et qui ont fait du mieux qu'ils pouvaient, malgré leur emploi du temps chargé, pour m'aiguiller et m'aider dans mes travaux. Merci également à Sylvain Lambinet d'avoir accepté d'être mon rapporteur.

Un grand merci à tous ceux qui ont aidé de près ou de loin à la réalisation de mes séquences. Didier Vinson, Garance Kim, Winnie Dhenin, Laure Haulet, François-Xavier Pedron, Maxime Gourdon, Benjamin Thelot et Jules Fernagut, qui ont réussi à jouer des scènes pas toujours très bien dirigées par l'ingénieur du son/réalisateur. Les collègues de son qui sont venus me donner un coup de main pour percher : Nicolas Vercambre, Baptise Mesange et Élise Guyonnet. Merci à Alexandre Delol qui a rendu possible ces tournages en me prêtant sa caméra, à Florine Bel pour m'avoir appris les rudiments du montage image ainsi que pour avoir étalonner une séquence qui en avait grandement besoin, et à Simon Bonanni pour son aide lors de l'écriture des séquences. Merci à Émilie Fretay, à la Base de Loisir de Cergy et à l'ENS Louis Lumière de m'avoir permis de trouver des décors pour mes séquences.

Et je remercie tout particulièrement mon chef opérateur, Maxime Sabin, sans qui rien n'aurait été possible et qui a su être patient et s'adapter à mon projet.

Merci à Simon Prieur pour ces deux semaines passées à mixer et à monter des structures en métal dans l'auditorium.

Je souhaiterais également remercier toutes les personnes qui ont pris part à mon expérience et qui m'ont permis de concrétiser mon projet.

Merci aux mixeurs qui ont bien voulu partager leur expérience et leurs impressions : Vincent Arnardi, Florent Lavallée, et Claude Gazeau.

Je tiens également à remercier Agnès Hominal qui m'a aidé sur de nombreux points administratifs lors de la préparation de mes tournages.

Bien évidemment, je remercie ma famille qui m'a soutenu lors de la réalisation de ce mémoire et lorsque je me sentais dépassé par les évènements.

Enfin, merci à tous les étudiants de la promotion 2016 de Louis Lumière.

Résumé

À l'heure où de nouveaux systèmes de reproduction du son au cinéma voient le jour (Dolby Atmos, Auro 3D et WFS), la question de la spatialisation des sons est plus que jamais d'actualité. L'utilisation notamment du mixage orienté objet semble offrir de nouvelles perspectives sur le placement et le déplacement des sources, ce qui relance le débat de la cohérence audiovisuelle spatiale des objets sonores. Mais les bandes sons au format 5.1 semblent aujourd'hui relativement codifiées, et il est notamment convenu de conserver la voix dans l'enceinte centrale.

Ce mémoire aura pour but d'étudier l'intérêt potentiel de la cohérence audiovisuelle spatiale en azimut des objets sonores et en particulier de la voix. Nous ferons d'abord l'état de la spatialisation des sons dans les bandes sonores actuelles et de la place de la voix au cinéma. Nous étudierons ensuite l'aspect psychoacoustique des relations entre la vue et l'ouïe et plus particulièrement "l'effet ventriloque". Enfin ces recherches mèneront à la mise en place d'un test perceptif visant à quantifier l'impact de la cohérence audiovisuelle spatiale en azimut. Ce test permettra notamment de voir l'importance de la nature de l'objet sonore (effet ou voix) et de sa vitesse à l'écran, sur le jugement des spectateurs.

Mots clés : *Cohérence audiovisuelle spatiale, spatialisation, voix, perception, effet ventriloque, déplacement, vitesse.*

Abstract

As new systems of sound reproduction in theater are launched (Dolby Atmos, Auro 3D and WFS), the question of sounds' spatialisation is more than ever topical. In particular, the use of object-oriented mixing appears to offer new perspectives for the placement and the movement of sound sources, which reopen the debate on the spatial audiovisual coherence for sound objects. But nowadays, 5.1 soundtracks appear to be relatively codified, for exemple the convention is to mix the voice in the central speaker.

The purpose of this dissertation will be to study the potential relevance of sound objects' spatial audiovisual coherence in azimuth, and paticularly of the voice. First we will make a state-of-the-art on sounds' spatialisation in modern soundtracks and on the role of voice in cinema. Then, we will be studying the psychoacoustic dimension of the relation between vision and audition, and in particular the "ventriloquist effect". Finally, this research will lead to the set up of an experiment aiming at quantify the impact of spatial audiovisual coherence in azimuth on audience's evaluation. This experiment will investigate the influence of the sound object's nature (effect or voice) and his speed across the cinema screen.

Mots clés : *Spatial audiovisual coherence, spatialisation, voice, perception, ventriloquist effect, movement, speed.*

Table des matières

Remerciements **1**

Résumé **3**

Introduction **7**

I État de l'art **9**

1 Le son spatialisé au cinéma **11**
1.1 Histoire de la spatialisation du son au cinéma 11
1.1.1 Les débuts du son au cinéma : prémices de la spatialisation 11
1.1.2 Les systèmes multicanaux 14
1.1.3 Le son numérique : la fin de l'encodage 18
1.2 Vers de nouveaux systèmes 21
1.2.1 L'Auro 3D 21
1.2.2 Le Dolby ATMOS 23
1.2.3 La WFS 26
1.2.4 Tendances de mixage actuelles, et évolutions possibles 27
1.3 La place de la voix au cinéma 29
1.3.1 Le cinéma qui parle 29
1.3.2 La voix spatialisée 32

2 Etat de l'art : perception image et son **37**
2.1 L'effet Ventriloque 37
2.1.1 Biais Intersensoriels 37
2.1.2 L'effet ventriloque en azimut 41
2.1.3 L'effet ventriloque en élévation 44
2.1.4 Discussion 45
2.2 Cohérence audiovisuelle spatiale des éléments sonores 45
2.2.1 Différentes études sur la cohérence audiovisuelle spatiale 46
2.2.2 Expérience V de la thèse de Etienne Hendrickx 50
2.3 Conclusion 53

II Contribution du mémoire **55**

3 Réalisation d'un test perceptif sur la cohérence audiovisuelle spatiale en azimut de la voix **57**

3.1 Choix des séquences et tournage 59
3.1.1 Écriture des séquences 59
3.1.2 Matériel 59
3.1.3 Méthodologie de tournage 60
3.2 Post-production des séquences et premières hypothèses 61
3.2.1 Montage son des séquences 61
3.2.2 Expérience de mixage 62
3.2.3 Hypothèses quant aux résultats de l'expérience 65
3.3 Description du test perceptif 66
3.3.1 L'installation technique 66
3.3.2 Le protocole 67
3.3.3 Descriptif des stimuli 69
3.4 Exploitation des résultats 72
3.4.1 Influence de la Nature de l'objet sonore 73
3.4.2 Corrélation entre la vitesse et la nature du stimuli 74
3.4.3 Influence de la vitesse de déplacement 76
3.4.4 Discussion avec les sujets 84
3.5 Conclusion 85

Conclusion générale **87**

Annexes **92**

Bibliographie **107**

Introduction

La notion d'écriture sonore au cinéma a toujours été très liée à la spatialisation des sons. L'utilisation du multicanal s'est imposée assez tôt au cinéma autant grâce à sa capacité à créer potentiellement une meilleure illusion de champ sonore naturel, que par les possibilités qu'il offrait sur l'écriture sonore. Mais si la spatialisation sonore a longtemps subi les contraintes dues directement aux systèmes de reproduction du son au cinéma, et notamment le matriçage du Dolby Stéréo, avec l'avènement du son numérique et des canaux discrets, l'écriture sonore peut à nouveau diriger les questions de spatialisation des sources sonores.

De plus, de nouveaux systèmes, à l'image du Dolby Atmos, de l'Auro 3D ou encore de IOSONO (utilisant la Wave Field Synthesis), proposent de meilleures performances sur le positionnement des sources sonores dans l'espace de la salle de cinéma. L'arrivée notamment de la notion de mixage objet renouvelle notre manière d'aborder le placement et le déplacement des sources sonores au cinéma.

Aujourd'hui, pour avoir une meilleure immersion, les mixeurs et les réalisateurs semblent assez ouverts aux déplacements de toutes sortes de sources sonores, que ce soit des ambiances, des bruitages ou des effets, et les spectateurs paraissent aujourd'hui "éduqués" à cette esthétique. Cependant, la voix semble avoir du mal à sortir des conventions établies par des années de cinéma monophonique. En effet, il est très rare que la voix sorte de l'enceinte centrale et ceci pour plusieurs raisons. Aussi, la voix reste dans le cinéma actuel l'axe central des films, et il est primordial que chaque spectateur puisse la comprendre et la percevoir sans artefacts. Mixer la voix au centre permet que chaque spectateur puisse percevoir la voix au centre de l'écran, et ce peu importe sa place dans la salle de cinéma. De plus, les études psycho-acoustiques ont montré l'existence d'un "effet ventriloque" permettant au cerveau de corriger les disparités spatiales entre un son et son correspondant visuel. Dans le cadre d'une projection cinématographique, cela veut dire que la cohérence audiovisuelle spatiale de la voix et du corps n'est pas nécessaire pour que le spectateur comprenne que les deux éléments sont associés.

Mais si le déplacement des sources sonores en tout genre semble pouvoir impacter positivement l'appréciation d'une séquence par les spectateurs, on peut légitimement se questionner sur la place de la voix. De plus, plusieurs films ont exploité la cohérence audiovisuelle

spatiale en azimut de la voix et ont su l'intégrer à l'écriture du film, comme *Gravity* d'Alfonso Cuaron (2013) et *Birdman* d'Alejandro G.Innaritu (2014). Les spectateurs n'ont pas eu l'air de s'en plaindre. Au contraire, ces films ont été largement appréciés notamment grâce à leur bande son : *Gravity* a remporté les oscars du meilleur montage son et du meilleur mixage en 2014 ; et *Birdman* a été nommé dans les deux mêmes catégories en 2015.

Ce mémoire tentera donc de répondre à une question simple : Dans quelles mesures la cohérence audiovisuelle spatiale en azimut de la voix au cinéma peut-elle améliorer l'expérience cinématographique faite par un spectateur ? Ainsi, nous essaierons également de voir les conditions qui pourraient gêner ou au contraire susciter l'intérêt de la cohérence azimutale des objets sonores pour le spectateur, que ce soient des voix ou bien des effets sonores.

Nous aborderons dans une première partie l'évolution de la spatialisation du son au cinéma, à travers les différents procédés de diffusion du son au cinéma. Nous décrirons également les nouveaux systèmes qui se commercialisent aujourd'hui et les possibilités qu'ils offrent. Nous étudierons dans cette même partie la place de la voix dans le cinéma actuel à la fois dans l'écriture, mais également dans son positionnement dans l'espace sonore.

Une deuxième partie s'intéressera à l'aspect psychoacoustique des relations entre l'image et le son. Aussi, nous étudierons les biais intersensoriels, c'est à dire l'influence de la perception de l'image sur la perception du son (et inversement). Cela sera également l'occasion d'étudier de plus près l'effet ventriloque. Nous verrons ensuite les études ayant porté sur la cohérence audiovisuelle spatiale.

Ces différentes recherches mèneront à l'établissement d'un test perceptif. En se basant sur des études précédentes, nous construirons un test permettant de mettre en évidence l'impact de la nature des objets sonores, que l'on séparera en deux catégories : la voix et les effets. Nous étudierons également la vitesse de ces sources sur les attentes des spectateurs en termes de spatialisation. Ce test permettra notamment de voir les limites de la cohérence audiovisuelle spatiale en étudiant l'impact du découpage d'une séquence. Aussi, nous aurons l'occasion de tester la cohérence audiovisuelle spatiale sur un des écueils de la spatialisation du son au cinéma, le cas du champ-contrechamp.

Première partie

État de l’art

Chapitre 1

Le son spatialisé au cinéma

Dans ce premier chapitre, nous étudierons d'abord brièvement l'histoire de la spatialisation du son dans le cinéma, depuis l'apparition du cinéma jusqu'aux nouveaux systèmes qui voient le jour aujourd'hui. Puis, nous verrons la place que la voix a pris dans le cinéma sonore.

1.1 Histoire de la spatialisation du son au cinéma

Avant d'aborder plus en profondeur la place de la voix au cinéma, nous ferons un rappel de l'évolution des systèmes de diffusion du son au cinéma, ainsi que l'évolution des manières d'aborder les questions de spatialisation du son.

1.1.1 Les débuts du son au cinéma : prémices de la spatialisation

Si le cinéma n'était pas considéré comme sonore, ou du moins "parlant", avant 1927 et la sortie de *Le chanteur de Jazz* de Alan Crossland, le son n'était pas pour autant absent des salles de cinéma.

Effectivement, le cinéma des débuts est dit "muet" et non "silencieux". C'est avant tout l'absence de voix accompagnant le corps des acteurs (pourtant bavards à l'écran) qui fait de lui un cinéma "muet". Cependant, il est habituel pour un film muet d'être accompagné par de la musique jouée en direct dans la salle de cinéma, par un groupe ou simplement un pianiste.

De plus, il est assez fréquent que les films soient également accompagnés par des bruiteurs directement présents dans la salle. Ceux-ci suivent une partition et soulignent les moments importants à l'image avec des bruitages. Il est de même courant de regarder un film muet avec un bonimenteur présent dans la salle. Il peut ainsi donner une voix soit aux personnages, soit aux informations sur l'histoire données par les cartons. C'est ce qu'Allain Boillat appellera la "voix vive" en opposition à la voix fixée sur enregistrement [1] .

1. Alain Boillat, *Du Bonimenteur à la voix-over*, Antipodes, 2007.

Alors que le cinéma muet se développe et trouve une manière de raconter les histoires, les technologies sonores évoluent de plus en plus. Si pour la musique, les cylindres et les disques se sont démocratisés et sont largement commercialisés, la compatibilité avec le cinéma se heurte aux problèmes de synchronisation image et son, à la piètre qualité des microphones de cette époque, ainsi qu'au défaut de puissance de rediffusion.

Cependant, rapidement des systèmes expérimentaux voient le jour proposant une synchronisation du son et de l'image. C'est notamment le cas de Edward H.Amet qui, en 1911, propose un système de diffusion basé sur la lecture de cylindre. Il propose de séparer le cylindre en plusieurs pistes et d'y enregistrer différents sons. L'idée est ensuite de relire les pistes et de les associer à des haut-parleurs de téléphones placés à différents endroits derrière l'écran. Afin, d'avoir un niveau plus important, il place un pavillon à la suite des haut-parleurs de téléphone.

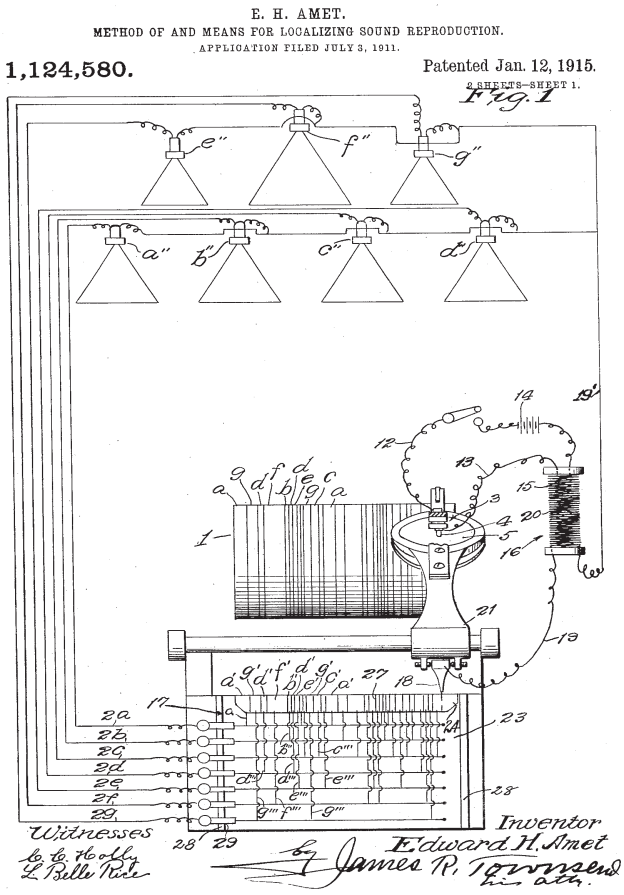


FIGURE 1.1 – *Schéma explicatif du système de diffusion de Edward H.Amet, avec ici 7 canaux de diffusion. (H.Amet, "Method of and means for localizing sound reproduction" Brevet technique déposé en 1911)*

Le but de ce système, présenté dans la figure 1.1, est à la fois de diffuser du son derrière l'écran de manière synchrone, mais également de pouvoir diffuser ce son au plus proche de l'endroit correspondant à l'image de la source à l'écran, grâce aux différents haut-parleurs. Si ce système ne sera finalement jamais utilisé, on peut noter qu'avant même la concrétisation du cinéma sonore, on recherche déjà une cohérence audiovisuelle spatiale entre les éléments sonores et leurs images.

Il faut attendre 1927 pour que les technologies de microphones, mais aussi de diffusion du son, permettent la sortie des premiers films sonores. À ce moment, deux méthodes s'opposent. *Le chanteur de Jazz* considéré comme le film pionnier du cinéma sonore, utilise une lecture depuis un disque synchronisé avec le projecteur. La même année, le film *L'heure suprême* de Frank Borzage utilise quant à lui l'inscription directement sur la pellicule à l'aide d'une piste "optique". On a alors une piste mono sur une pellicule 16mm. C'est cette méthode qui se démocratisera et sera utilisée par la suite.

Il faudra un temps d'adaptation à ce nouveau cinéma, autant pour les spectateurs que pour les cinéastes. Le cinéma doit se reconstruire et trouver des manières différentes de raconter les histoires. L'apport du son oblige à redéfinir la temporalité des films, certaines astuces du muet ne sont plus possibles, comme de lire des plans à l'envers par exemple. De même, les spectateurs sont décontenancés, à la fois par les films proposés, mais aussi par le système de diffusion. Mais l'habitude aidant, les spectateurs ont accepté ce mode de diffusion.

Dans ce cinéma monophonique des débuts, on a ce que Claude Baiblé appelle un "effet de hublot"[2], c'est-à-dire que toutes les informations passent par un seul point de diffusion. Et c'est alors le cerveau qui fait lui même le travail de localisation des sons à partir des différents indices qui lui sont fournis. Aussi, l'image permet à cette localisation imaginaire d'avoir lieu, en utilisant le montage, ainsi que les regards des personnages.

Mais si à cette époque et jusqu'au début des années 1940, la norme est au film monophonique, différentes tentatives de multicanal sont à répertorier. Tout d'abord, de nombreux cinémas sont déjà équipés d'enceintes placées dans la salle et non derrière l'écran, et si la bande son du film reste en mono, il n'est pas rare que durant la séance certains passages soient diffusés dans les enceintes "surrounds". En effet, un technicien est présent lors de la diffusion et utilise des commutateurs afin d'envoyer le son du film tour à tour dans les enceintes derrière l'écran, ou dans les enceintes arrières. Il connaît très bien le film et sait à quels moments il devra faire les changements. Par exemple, les passages musicaux sont souvent diffusés dans la salle plutôt que derrière l'écran. Il existe également des systèmes avec plusieurs enceintes derrière l'écran, et le technicien commute alors entre les différentes enceintes afin d'avoir des placements de son plus cohérents à l'image.

2. Claude Baiblé, "L'image frontale, le son spatial" dans *Cinéma et dernières technologies* dirigé par Frank Beau, Philippe Dubois et Gerard Leblanc, INA et De Boeck et Larcier, 1998, p.235

Ensuite, comment oublier l'exemple célèbre du film d'Abel Gance, *Napoléon* ? On parle ici de la version de 1935 revisitant la version initiale de 1927 alors muette, et qui sera également appelée *Napoléon vu et entendu par Abel Gance*. Ce film est le premier à réellement utiliser une bande son en multicanal. Présenté avec trois projecteurs dirigés vers trois écrans adjacents, le film d'Abel Gance est sans précédent. Pour la diffusion du son, une enceinte est disposée derrière chacun des trois écrans, et on a un système se rapprochant d'un LCR (Left Centre Right) comme on les connaît aujourd'hui. De plus, 32 haut-parleurs sont disposés dans la salle à divers emplacements. Ceux-ci sont synchronisés avec les projecteurs afin de lancer la lecture de sons spécifiques durant le film à des moments prédéterminés [3]. Ce film propose alors une nouvelle manière de penser le son au cinéma. Mais il n'est projeté que dans une salle à Paris (la salle Paramount) ; son installation beaucoup trop complexe et onéreuse ne s'étendra pas plus loin que son film, et il faudra attendre encore un peu avant de voir le multicanal se démocratiser.

FIGURE 1.2 – *Projection de "Napoléon vu et entendu par Abel Gance", avec ses trois écrans.*

1.1.2 Les systèmes multicanaux

C'est en 1940 que le multicanal fait réellement son apparition dans les salles de cinéma. Deux systèmes apparaissent. Le premier, le "Vitasound" présente une version améliorée des systèmes à commutation existant déjà. Ici, la commutation entre l'enceinte frontale et les enceintes arrières n'est plus contrôlée par un technicien dans la salle, mais directement par une piste de contrôle. Deux pistes sont donc présentes sur la pellicule, une piste contenant le son monophonique et une piste de commande automatisant les passages de l'avant vers l'arrière.

3. On a par exemple des sons de fusils, de chevaux ou même des voix de soldats interpellant le public.

Mais la même année, un système plus innovant apparaît : le "Fantasound". Mis au point spécialement pour le film *Fantasia*[4] de Disney, c'est le premier système proposant du multicanal (et non de la commutation d'une seule piste). Ce système possède alors trois enceintes placées derrière l'écran de projection et un canal placé derrière les spectateurs. À cette époque on ne sait pas placer plus de deux pistes optiques sur la pellicule, il faut donc un deuxième projecteur synchronisé avec le premier, servant exclusivement pour diffuser le son. Cette installation est excessivement coûteuse et seulement six cinémas aux États-Unis s'équipent avec le système Fantasound ; les autres diffusent le film en mono. Le Fantasound est donc arrêté très rapidement et n'aura été utilisé que pour un seul film. Il est néanmoins le précurseur de la configuration qu'utiliseront plus tard de nombreux systèmes multicanaux, à savoir trois canaux frontaux et un canal arrière.

Il faut ensuite attendre les années 50 pour voir de nouveaux systèmes arriver, et ce dû à la fois à la meilleure qualité des enceintes et à l'apparition de la bande magnétique plus facile à utiliser et avec une meilleure qualité sonore que la piste optique. Plusieurs systèmes avec bandes magnétiques arrivent sur le marché, la plupart du temps associés à un système de prise de vue et de projection particulier. Ces nouvelles technologies permettent d'effectuer pour la première fois des panoramiques de sources sonores, notamment grâce au meilleur rapport signal sur bruit ; de même le canal arrière peut être allumé en permanence, alors qu'il était auparavant coupé (par une piste de contrôle) lorsqu'il n'était pas utilisé.

Ainsi, en 1952, le *Cinerama* est présenté. C'est un système utilisant trois projecteurs et donc utilisant des immenses écrans incurvés. Un quatrième projecteur est associé pour le son. Sept canaux sonores sont employés avec cinq enceintes frontales, avec pour la première fois l'utilisation de canaux Inter-Gauche et Inter-Droit. Deux enceintes sont utilisées pour les surrounds. L'installation étant extrêmement coûteuse[5], très peu de salles s'équipent pour ce procédé. Le système est utilisé jusque dans le début des années 60, avec à son actif uniquement 10 films dont seulement deux sont des fictions[6]. On peut noter que l'URSS propose dès 1957 un système similaire : le Kinopanorama. Si la projection de l'image se fait également via trois projecteurs, il propose, contrairement à son homologue américain, 9 pistes magnétiques, avec l'ajout d'un canal centre-arrière et d'un canal au dessus des spectateurs.

4. Coréalisé par : J. Algar,S. Armstrong, F. Beebe, N. Ferguson, J. Handley, T. Hee, W. Jackson, H. Luske, B. Roberts, P. Satterfield.

5. Les tournages sont également plus compliqués à mettre en place avec trois caméras nécessaires, et l'utilisation également de six microphones.

6. Les autres films sont des documentaires. On se rapproche de ce qu'on peut voir par exemple à *La Géode* de la Cité des Sciences à Paris.

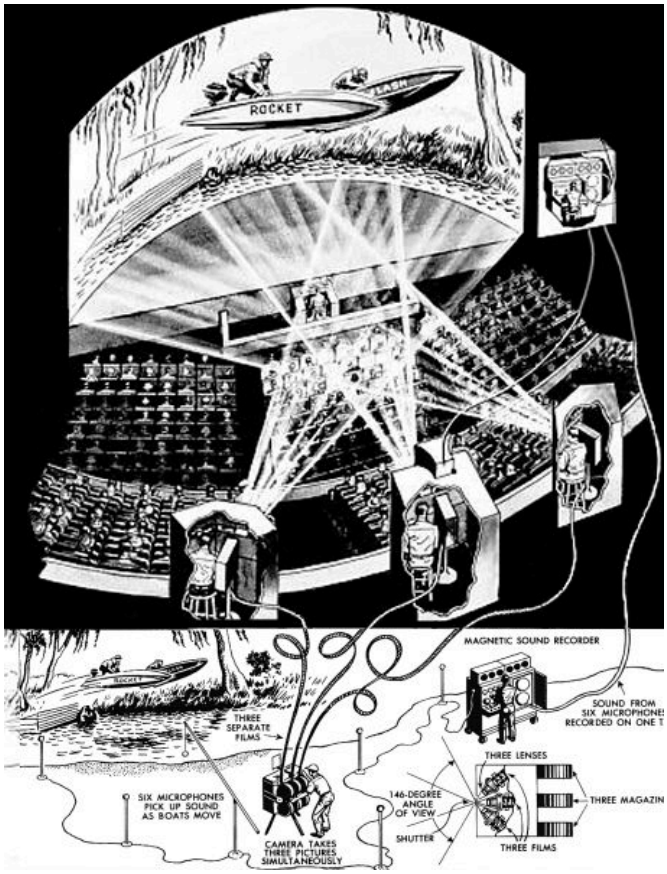


FIGURE 1.3 – *Présentation du système Cinerama, avec ses trois cabines de projections et sa cabine pour la lecture du son.*

Les systèmes Todd AO et SuperPanavision 70, qui utilisent tous les deux des pellicules 70mm, emploient un dispositif similaire pour le son. Pour couvrir la taille importante de la projection, six pistes magnétiques sont présentes sur la pellicule, et viennent alimenter cinq enceintes derrière l'écran et une enceinte arrière. Cependant, ces deux procédés impliquent de nombreux changements pour les exploitants, notamment car la norme de diffusion à l'époque est le 35mm. C'est une des raisons pour laquelle un autre procédé s'est imposé, le Cinemascope.

Contrairement à ses concurrents, le cinémascope utilisent une technique d'anamorphose qui lui permet de placer une plus grande image sur une pellicule 35mm. Ainsi, les exploitants n'ont qu'à s'équiper de nouvelles lentilles et non de nouveaux projecteurs. Pour ce qui est du son, le cinémascope propose une solution à quatre canaux, similaire à la configuration du Fantasound, avec un LCR frontal et un canal de surround. La Fox, qui commercialise le cinemascope, oblige d'abord les salles qui veulent passer leurs films à s'équiper d'un système son adéquat, et à la fin des années 50, plus de 10000 cinémas sont équipés pour diffuser des films en cinémascope. Cependant, les spectateurs ont du mal à se faire aux premières expérimentations du son surround et au déplacement des sons. La Fox change de politique et

tous ses films sont alors disponibles en 4.0 mais également avec un mixage mono sur piste optique. Les cinéastes préfèrent donc limiter leur utilisation du surround, ou encore de prise de son stéréo, et se limitent à des sons monophoniques, parfois pannés sur les canaux frontaux, afin d'assurer la retrocompatibilité monophonique.

Jusqu'au début des années 70, on ne voit pas apparaître de nouveaux procédés, et le cinemascope est le système principal, mais la création cinématographique est revenu à un cinéma principalement monophonique. C'est à partir de 1974 que l'on voit naître de nouvelles tentatives. Ainsi, c'est cette année-là que le système Sensurround est utilisé pour la première fois sur le film *Tremblement de terre* (de Mark Robson). Ce procédé emploie des caissons de basses pour avoir plus de sensation d'immersion. Le sensuround n'aura pas de vrai succès et s'arrêtera en 1981 après seulement 5 films.

L'année suivante, deux systèmes sont commercialisés. Les deux surfent sur la vague de succès que rencontre le format quadriphonique en musique qui offre quatre canaux sur deux pistes magnétique grâce au matriçage. Le système Quintaphonic reprend alors complètement le principe de la quadriphonie en lui ajoutant une troisième piste représentant un Centre non matricé, il présente donc des arrières stéréo.

Mais face à lui, le Dolby stéréo, créé la même année, propose un système économiquement beaucoup plus avantageux pour les studios et pour les exploitants de salles. En effet, Dolby profite de sa technologie de réduction de bruit Dolby NR-C pour pouvoir utiliser des pistes optiques qui coûtent beaucoup moins cher à faire que les pistes magnétiques. De plus, contrairement au quintaphonic, il reste sur une configuration à quatre canaux LCRS (Left Centre Right Surround) comme celle du cinemascope. Tout le monde y trouve son compte : la piste optique coûte moins cher aux studios et les cinémas déjà équipés pour le cinémascope n'ont qu'à acquérir un décodeur Dolby stéréo. De plus, la version 70mm du Dolby stéréo utilise les enceintes Inter-Gauche et Inter-Droite des systèmes précédents comme renfort de basse. Pour la première fois, un format propose une solution avec très peu de changements à prévoir de la part des salles de cinéma. De plus, le système Dolby stéréo est totalement compatible avec les salles encore en mono. La sortie en 1977 de *Star Wars* (de George Lucas) concrétise l'utilisation du Dolby Stéréo.

Si le Dolby stéréo s'impose massivement, il n'en est pas pour le moins très controversé, et il est notamment vu par les réalisateurs ou les mixeurs comme un retour en arrière. En effet, le dolby stéréo repose entièrement sur la méthode de matriçage, qui permet d'avoir les quatre canaux LCRS sur deux pistes optiques. Mais c'est ce même matriçage qui devient une énorme contrainte pour les mixeurs. En effet, contrairement aux systèmes discrets des années 50, le matriçage ne permet pas aux ingénieurs du son d'avoir un contrôle total sur le rendu sonore. Le matriçage empêche de mettre des sons importants pour la narration dans le canal arrière,

de même l'utilisation de la phase pour matricer le canal arrière rend compliqué l'intégration de contenu stéréo. Kerins [7] souligne que, dès ses débuts, plusieurs films ont contourné les codes du Dolby Stéréo en ajoutant par exemple un subwoofer (*Rencontres du troisième type*, Steven Spielberg, 1977), ou des arrières stéréo (Dans *Superman* de Richard Donner en 1978) ou avec un système proche du quintaphonic dans *Apocalypse Now* de Francis F. Coppola (1979) ; ces films sont néanmoins sortis avec l'appellation Dolby Stéréo, cela participant probablement à augmenter sa suprématie sur les systèmes existants.

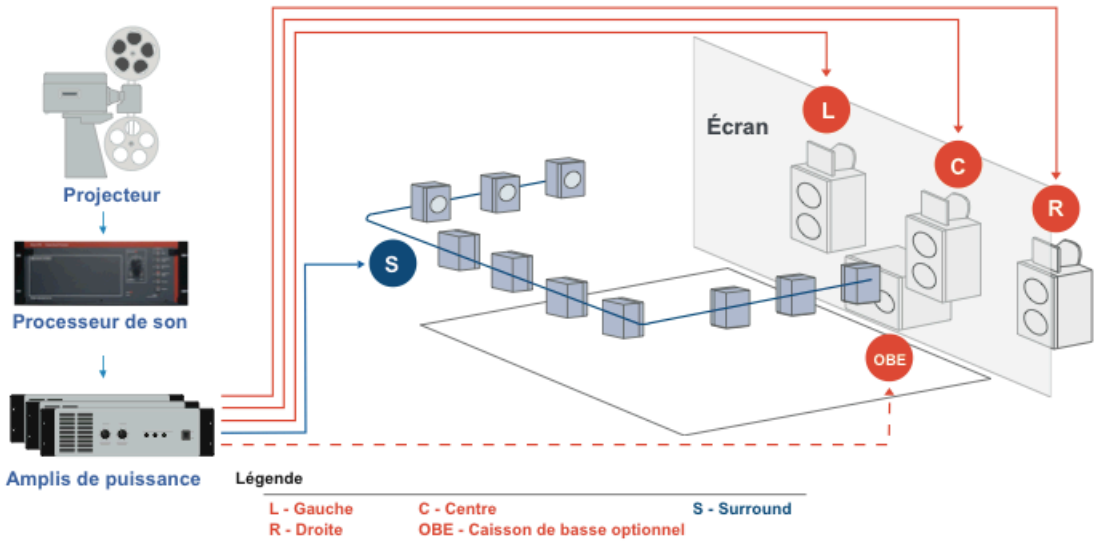


FIGURE 1.4 – *Configuration pour une salle au format Dolby Stéréo (Dolby, 2012)*

Mais les problèmes liés au matriçage persistent. De plus, les arrières ne sont pas pleines bandes contrairement aux enceintes avant. Dolby mise tout sur la compatibilité mono de son système, qui de surcroît à tout pour dissuader les ingénieurs du son d'utiliser trop de spatialisation. Ainsi, Stanley Kubrick décidera de réaliser tous ces films en monophonie afin de garder le contrôle sur le son de ses films et ne pas dépendre du bon vouloir du décodage Dolby. La spatialisation des sons au cinéma est donc encore et toujours contrainte par des contingences techniques et économiques.

1.1.3 Le son numérique : la fin de l'encodage

La démocratisation du Compact Disc en musique et les évolutions du son numérique aidant, la SMPTE (Society of Motion Picture and Television Engineer) commence à travailler en 1987 sur un nouveau standard pour le son numérique au cinéma. Pour ce nouveau standard il y a deux objectifs :

- La disparition du matriçage, uniquement des canaux discrets.
- Une qualité sonore au moins égale à la qualité du CD audio.

7. *Beyond Dolby (stereo)*, Indiana University Press, 2011

Le choix est alors fait d'adopter un système proche du quintaphonic : le 5.1. Ce système présente trois canaux frontaux et deux canaux arrières [8], mais utilise également un canal supplémentaire réservé aux fréquences basses, c'est le LFE (Low Frequency Effect).

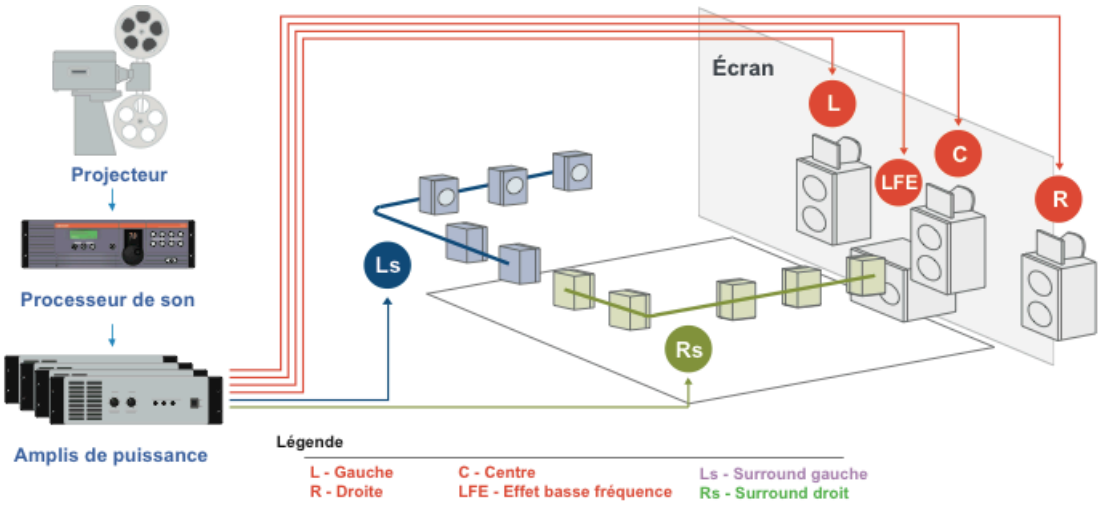


FIGURE 1.5 – *Configuration pour une salle au format 5.1 (ici système Dolby SR-D)(Dolby, 2012)*

Le premier système, le CDS (Cinema Digital Sound) sort en 1990 avec *Dick Tracy* de Warren Beatty. Cependant, moins de dix films sortiront sous ce format. Effectivement le système possède un défaut : il ne prévoit pas de pistes analogique et ne permet donc pas la rétrocompatibilité.

Trois systèmes vont ensuite coexister pendant plus de 10 ans. Il s'agit du Dolby Sr-D [9] créé en 1992 pour *Batman Returns*, du DTS (Digital Theater Systems) et du SDDS (Sony Dynamic Digital Sound) créés tous deux en 1993 respectivement pour *Jurassik Park* et *Last Action Hero* [10] . Les trois proposent une piste optique Dolby Stéréo pour la retrocompatibilité.

Le Dolby Digital et le DTS proposent tous les deux des mixages en 5.1, mais par des techniques différentes. Alors que Dolby compresse le son en AC-3 (compression 1/12) et place l'information directement sur la pellicule, DTS utilise un lecteur CD externe synchronisé avec une piste de Time Code sur la pellicule, permettant ainsi une meilleure qualité sonore (compression 1/3). Le système DTS coûte également moins cher à installer.

Le SDDS, contrairement à ses concurrents, propose une configuration en 7.1 avec des Inter-Gauche et Inter-Droit. Si l'installation coûte plus cher, le système de Sony mise sur la compatibilité avec les anciennes salles équipées pour les projections en 70mm.

8. Ici les cinq canaux sont pleine bande.
9. Dolby Spectral Reduction-Digital qui deviendra plus tard le Dolby Digital.
10. *Batman Returns* de Tim Burton, *Jurassik Park* de Steven Spielberg et *Last Action Hero* de John McTiernan.

Les trois systèmes sont chacun soutenus par des studios différents, Universal soutient DTS, Columbia le SDDS et Dolby est appuyé par Paramount et Disney. Un accord pour un standard semble donc compliqué. Cependant, les trois systèmes utilisent tous un endroits différents sur la pellicule, on peut donc imprimer les trois versions du mixage sur une seule copie, permettant aux exploitants d'être assurés de pouvoir passer le film. Les années 1990 correspondent à une période de multiplication du nombre de copies de film possible et du nombre de cinémas (qui ouvrent alors équipés pour le son numérique). Tous ces paramètres permettront au son 5.1 de s'imposer.

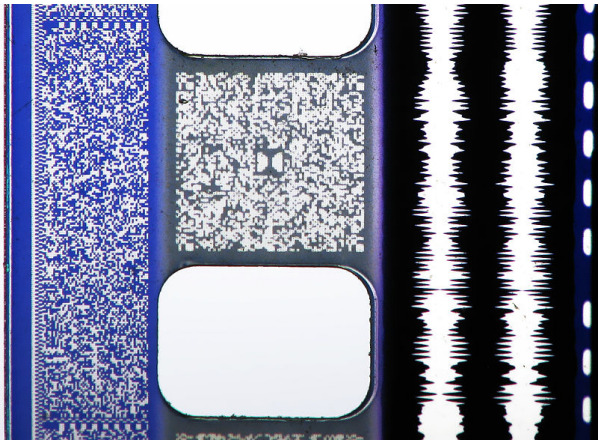


FIGURE 1.6 – *Pistes sonores sur une pellicule : Le SDDS placé à l'exterieur des perforations, le Dolby Digital entre les perforations, la version analogique Dolby Stéréo à droite des perforations, et le Time Code pour le DTS sur le bord de l'image (ici à droite).*

Cette expansion du 5.1 dans les cinémas est également corrélée à une expansion du nombre de home-cinémas et à l'exploitation télévisuelle des films. La majorité des profits d'un film vient de la vente de DVD et de la diffusion télé. Or, si la télé s'ouvre de plus en plus au multicanal (avec des téléfilms mixés en 5.1), la majorité des films seront regardés en stéréo classique. À nouveau se pose donc la question de l'utilisation des arrières dans la narration, puisqu'un mixdown en stéréo rendra forcément tous les sons de manière frontale. Si Dolby et DTS sont présent sur le marché du home-cinéma, SDDS n'est pas de la partie, préférant se limiter à la salle de cinéma. De plus, sa configuration à 5 enceintes frontale serait trop onéreuse et démesurée pour les petits écrans des home-cinémas.

Si DTS et Dolby font à peu près jeu égal en terme de nombre de salles équipées et de films à ce format (avec une légère avance pour Dolby), SDDS semble en retard, et arrête dès le début des années 2000 de vendre son système aux exploitants, perdant peu à peu du terrain sur ses concurrents. Dans le même temps, Dolby et DTS améliorent également leur système : pour *Star Wars : La menace fantôme* de G. Lucas (1999), Dolby rajoute un canal arrière centrale BS (Back Surround) matricé, c'est le Dolby Surround EX. DTS suivra avec son DTS ES avec un Back surround non matricé.

L'arrivée du cinéma numérique à la fin des années 2000 permet d'intégrer jusqu'à 16 canaux par DCP (Digital Cinema Package) et permet également de pouvoir diffuser des bandes son sans compression de données, comme c'était le cas jusqu'à présent. C'est dans ce contexte que Dolby présente son système Dolby surround 7.1 pour le film *Toy Story 3* (de Lee Unkrich, 2010). Contrairement au SDDS, ici on conserve le LCR et on sépare les surrounds en 4 canaux, les deux canaux latéraux Lss et Rss (Left/Right Side Surround) et les deux canaux Lrs et Rrs (Left/Right Rear Surround).

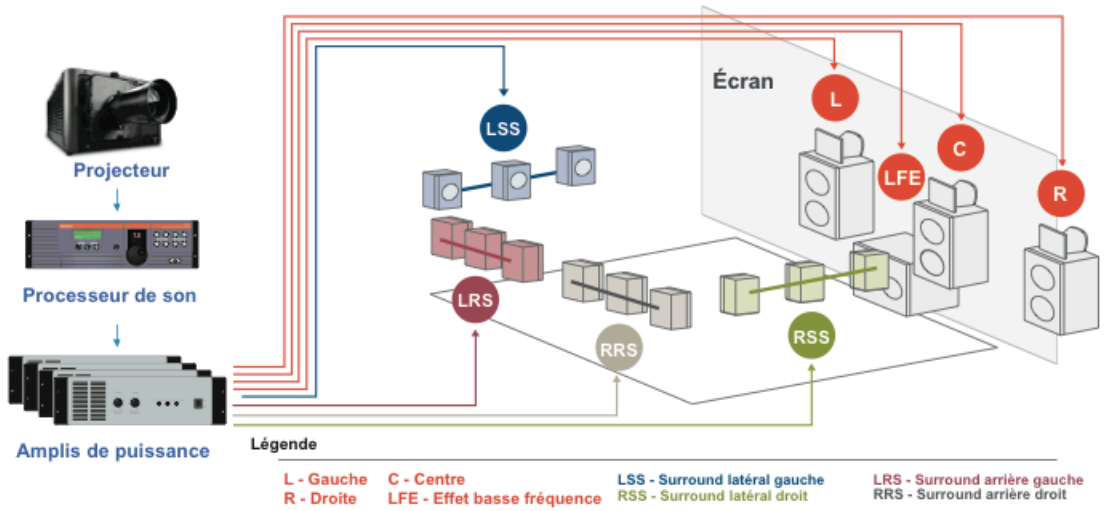


FIGURE 1.7 – *Configuration pour une salle au format Dolby Surround 7.1 (Dolby, 2012)*

1.2 Vers de nouveaux systèmes

Le format 5.1 reste aujourd'hui le format le plus largement répandu et utilisé, et les réalisateurs et les ingénieurs du son ont su apprivoiser ce format pour faire passer leurs intentions. Cependant, de la même manière que le cinéma évolue sur les formats d'image avec la démocratisation de la 3D, l'IMAX ou encore le 4K, de nouveaux systèmes sonores apparaissent, et offrent de nouvelles perspectives pour les bandes son de film.

1.2.1 L'Auro 3D

L'Auro 3D est un format développé dès 2005 par la société Auro Technologie, et appartenant aujourd'hui au groupe Barco. L'idée de l'Auro 3D est d'étendre le système actuel du 5.1 à la dimension verticale. Auro 3D propose donc un système à trois niveaux :

- Un niveau "bas" composé d'une configuration 5.1 classique.

- Un niveau "haut" composé d'un 5.0 situé environ 30 degrés au dessus du niveau classique.

- Un canal "top" aussi appelé "voice of god" situé au dessus des spectateurs.

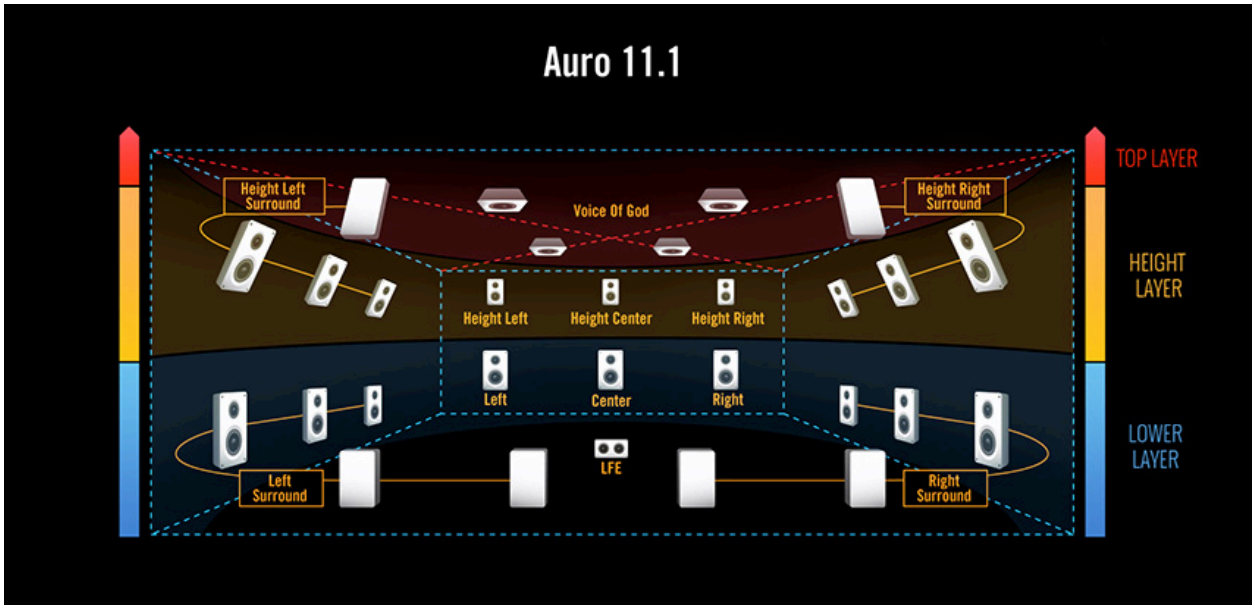


FIGURE 1.8 – *Configuration Auro 3D 11.1 (Auro-3D, 2006)*

Une version de l'auro 3D en 13.1 rajoute les surrounds latéraux du 7.1 de Dolby. Le système ainsi conçu permet de se baser sur les installations existantes dans les salles et de les améliorer en rajoutant peu d'enceintes. De plus, la retrocompatibilité avec un 5.1 est assez aisée. Les mécanismes de placement des sources restent les mêmes que dans les systèmes précédents. Contrairement à ses concurrents, l'Auro 3D est le seul à proposer pour l'instant des systèmes de prise de son adaptés à l'Auro 3D, avec des ensembles à 9, 10 ou 11 microphones.

Aujourd'hui, l'Auro 3D est présent dans plusieurs centaines de salles à travers le monde, et a été utilisé pour de très nombreux films depuis *Red Tails* de Anthony Hemingway en 2012 [11], avec notamment *Les Gardiens de la galaxies* (James Gunn 2012), *Les Croods* (C. Sanders et K. DeMicco 2013) ou encore *Le Monde Fantastique d'OZ* (Sam Raimi 2013) [12]. En France, on ne compte pour l'instant que trois salles équipées [13], et aucune production française n'a utilisé cette technologie.

Comme Dolby et DTS avant lui, l'Auro 3D est présent sur le marché du home-cinéma avec des versions plus légères de ces systèmes en 9.1 ou 10.1.

22.2 NHK

On peut noter que l'Auro 3D se rapproche également d'un système développé par la NHK (compagnie de télédiffusion Japonaise), le 22.2, conçu initialement pour la télévision en

11. Premier film au format Auro 3D 11.1.
12. On compte notamment de nombreuses production asiatique notamment à Bollywood.
13. La salle 16 de l'UGC la Défense, le "Cinelilas" aux Lilas, et un cinéma Pathé à Caen.

Ultra HD. Ce système exploite lui aussi la dimension verticale des bandes sons et fonctionne également sur 3 niveaux :

- Un niveau "bas" composé de 3 enceintes frontales situées au bas de l'écran.

- Un niveau "moyen" composé d'un système à 10 enceintes : 5 derrière l'écran, 2 sur les côtés, 3 derrière les spectateurs.

- Un niveau "haut" composé de 9 enceintes : 3 enceintes frontales, 2 sur les côtés, 3 derrières les spectateurs et un canal au dessus du public.

- S'y ajoutent deux LFE.

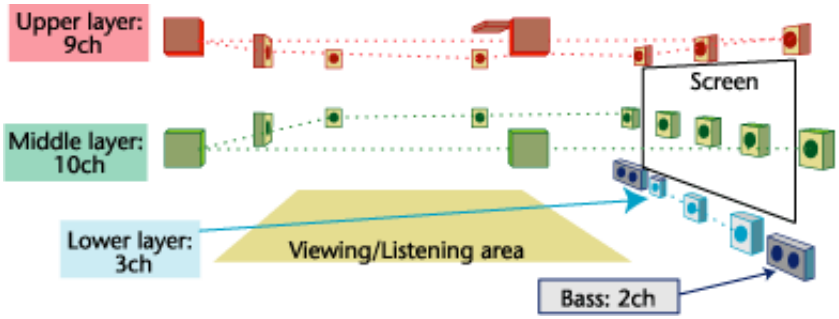


FIGURE 1.9 – *Configuration du 22.2 NHK*

Ce système, bien qu'il ait été développé dès le milieux des années 2000, n'est pas encore implanté. NHK souhaite l'intégrer à la diffusion télé au Japon pour 2016.

1.2.2 Le Dolby ATMOS

Parmi les nouveaux systèmes, le Dolby Atmos présenté en 2012 est celui qui est aujourd'hui le mieux implanté dans le paysage cinématographique, fort de l'empire créé par Dolby depuis plusieurs dizaines d'années. Le premier film mixé dans ce format est *Rebelle* des studios Disney sorti en 2012, mais c'est le film *Gravity* d'Alphonso Cuaron (2013) qui rend le système populaire. Aujourd'hui, avec plus de 160 films mixés en Atmos [14] depuis 2012 et plus de 2000 salles équipées dans le monde, le Dolby Atmos a pris une certaine avance.

En France on compte une trentaine de salles équipées [15]. Mais beaucoup des films mixés en Atmos ne sont diffusés en France que sous leur version 5.1/7.1. De plus, la production française est pour l'instant assez timide pour utiliser ce format, malgré l'implantation du système Atmos dans plusieurs auditoriums de mixage parisien. On compte très peu de film français en Atmos : *Taken 2* et 3 (Olivier Megaton 2012,2014), *Lucy* (Luc Besson, 2014), *En solitaire* (Christophe Offenstein, 2013), *Pourquoi j'ai pas mangé mon père* (Jamel Debbouze 2015) et

14. Beaucoup de ces films sont également mixé en Auro 3D.
15. La première salle ayant été équipée étant au Pathé Wepler à Paris.

plus récemment *Les Saisons* (Jacques Perrin, 2015). De plus, plusieurs d'entre eux n'ont pas vu leur version en Atmos mixée en France.

Principe et implantation dans la salle

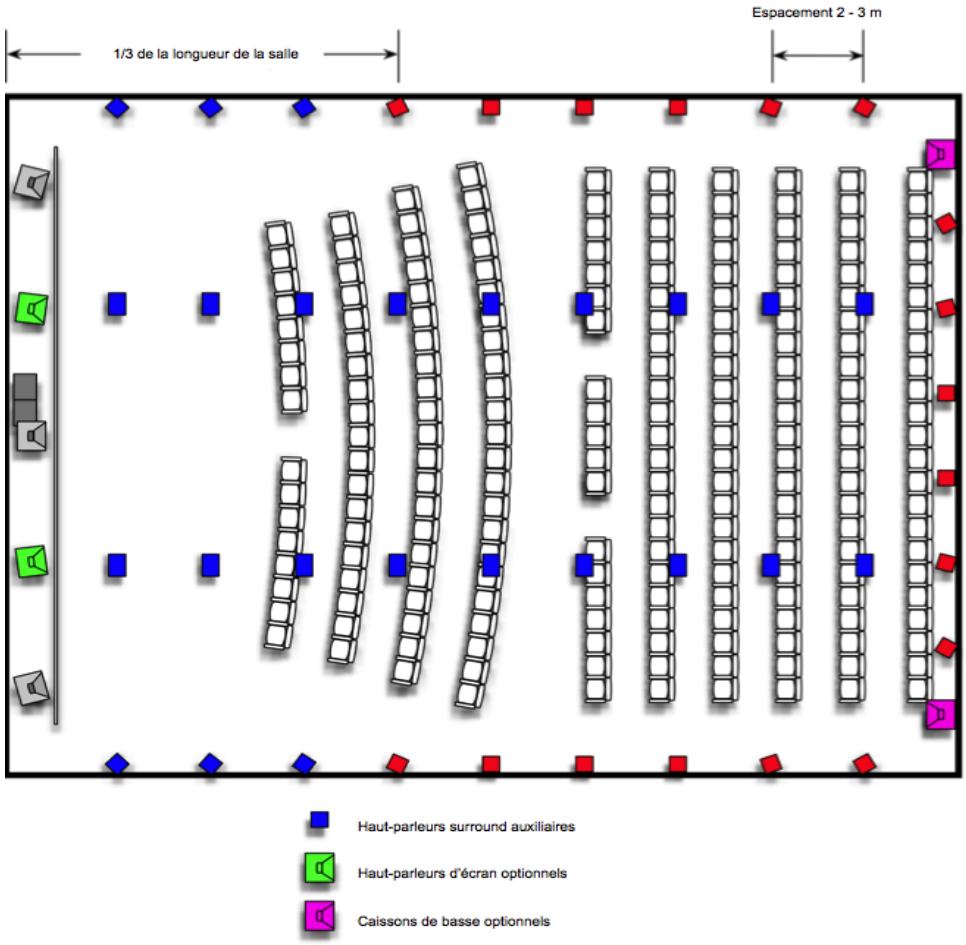


FIGURE 1.10 – *Implantation d'un système Atmos dans une salle de cinéma. En rouge on voit les haut-parleurs surround déjà existant, les haut-parleurs surround auxiliaires sont représentés en bleu, tandis que les subwoofer additionnels sont en rose. (Dolby, 2012)*

Dolby Atmos propose une nouvelle configuration pour le son dit "3D" au cinéma. Comme son concurrent direct l'Auro 3D, le Dolby Atmos fait entrer la notion de verticalité dans le mixage, mais le plus important dans ce système est la notion de mixage orienté objet, qui s'oppose au mixage traditionnel qui est basé sur les canaux. Le principe du mixage orienté objet est de séparer l'information audio des informations spatiales qui lui sont associées.

Le dolby atmos propose d'associer à la fois un flux mixé traditionnellement en 9.1 (ici on ajoute au 7.1 deux canaux au dessus du public) c'est ce qu'on appellera le "BED", et un ensemble d'objets sonores qui peuvent être placés indépendamment selon les informations

de spatialisation associées. De plus, ces objets pourront être diffusés sur n'importe quelle enceinte du système individuellement et on n'a plus la contrainte liée à la diffusion par canal, où plusieurs enceintes spatialement séparées diffusent le même son (typiquement les enceintes de surround). Le flux Atmos est donc composé de 128 pistes (dont 10 sont réservées au BED), ce qui laisse la possibilité d'avoir 118 objets sonores.

Lors du mixage, le mixeur décidera quels sons seront mixés traditionnellement dans le 9.1, et quels sons seront mixés comme des objets sonores. Le mixage orienté objet présente plusieurs intérêts : le mixeur pourra placer très précisément des sons, puisque chaque enceinte peut être gérée independamment. De plus, le fait de séparer le contenu audio des informations de spatialisation permettra une bonne transportabilité d'une salle à une autre. En effet, théoriquement le rendu de l'espace ne dépendra pas de la salle de restitution, c'est le processeur de rendu du Dolby Atmos, le RMU (Rendering Mastering Unit), qui adaptera la restitution à la configuration d'enceintes de la salle. Il n'est donc pas obligatoire que chaque salle soit équipée exactement de la même manière.

FIGURE 1.11 – *Principe du mixage orienté objet avec ces beds et ses objets sonores (Dolby, 2012).*

Aujourd'hui, plusieurs solutions existent pour travailler en Atmos : on peut utiliser des plug-ins permettant de panner en atmos, ou on peut mixer sur une console DFC Gemini de Neve, qui comprend un panner 3D pour l'Atmos.

Comme on peut le voir sur la figure 1.10, en plus de l'ajout des deux rangées d'enceintes au plafond, on remarque la présence d'enceintes entre l'écran et le premier tiers de la salle [16]. Celles-ci ne peuvent être utilisées que pour le placement des objets. Avec ces enceintes et le mixage objet, on peut donc obtenir un déplacement régulier du centre de l'écran jusqu'à l'arrière de la salle, ce qui n'était pas envisageable avec les systèmes basés sur les canaux. Le mixage orienté objet ouvre donc la voie à de nouvelles manières d'appréhender le placement des sources sonores au cinéma.

16. Où traditionnellement on place les premières enceintes surround

1.2.3 La WFS

La WFS (Wave Field Synthesis) a été développée à la fin des années 1980. Elle se base sur le principe de Huygens, qui dit que 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), est de faire une approximation du front d'onde d'origine en utilisant des haut-parleurs servant de sources secondaires. Ainsi, en superposant les ondes secondaires générées par ces haut-parleurs, on obtiendra un champ sonore proche d'une restitution physique et naturel. Contrairement aux autres systèmes qui ne font que simuler de la latéralisation et des espaces par des illusions auditives et autres images fantômes, la WFS tente de reproduire un champ acoustique naturel. Comme l'Atmos, la WFS a une approche du mixage orientée objet.

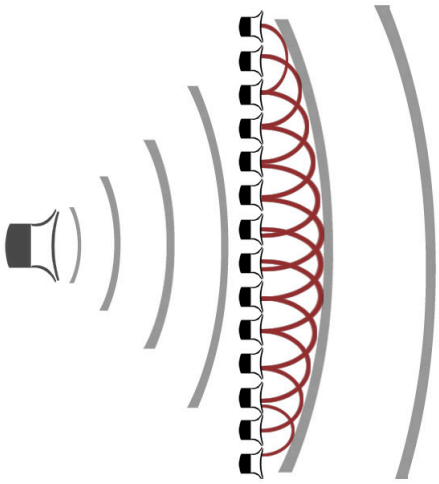


FIGURE 1.12 – *Illustration du principe de Huygens dans le cadre d'une restitution en WFS.*

Depuis le début des années 2000 la WFS est envisagée comme un système de reproduction pour le cinéma. Dès 2003, le cinéma de Illmenau en Allemagne est équipé de 192 haut-parleurs. Suivront des installations dans quelques salles aux États-Unis. Le projet est principalement porté par l'entreprise allemande IOSONO qui, depuis 2014, fait partie du groupe BARCO, également distributeur des systèmes Auro 3D. Le premier film bénéficiant de la technologie de IOSONO est *Les immortels* en 2011 (réalisé par Tarsem Singh).

Pour le cinéma, la WFS présente de nombreux atouts : une meilleure restitution de l'espace, une précision de localisation quasi-similaire à la localisation d'une source réelle, une bonne restitution de la profondeur et un suivi des sources plus naturel. Mais le principal atout réside dans l'absence théorique de "Sweet Spot". En effet, tous les systèmes existants ne fournissent un champ sonore "cohérent" que pour un nombre limité de place dans la salle de cinéma : le "sweet spot". Le mixeur étant lui même situé à cette position, il ne mixe "que" pour

les gens situés dans cette petite zone de la salle [17]. La WFS en restituant un champ sonore proche d'une restitution physique des sources, augmente considérablement la zone d'écoute. Peu importe l'emplacement des spectateurs dans cette zone d'écoute, ils percevront tous les sources provenant du même endroit. Théoriquement cette zone devrait être infinie, mais la WFS, de par l'utilisation d'un nombre restreint de haut-parleurs, n'est qu'une approximation et la zone, bien qu'elle soit considérablement plus grande que pour les systèmes classiques, reste limitée. On peut noter également que des essais fructueux ont été effectué pour étendre la WFS sur le plan vertical.

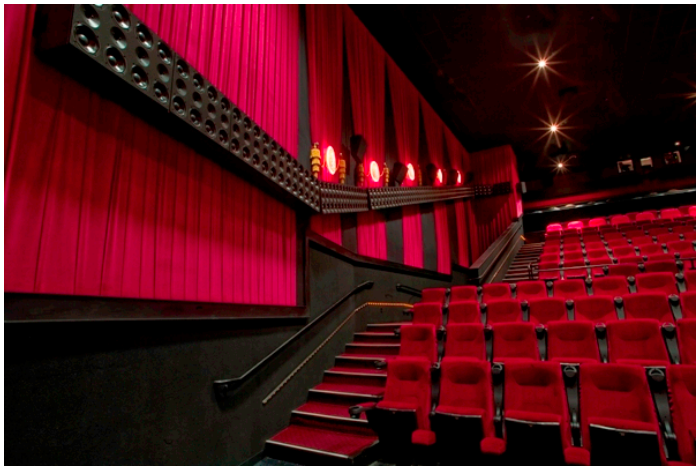


FIGURE 1.13 – *La salle du Mann Chinese Theater à Los Angeles équipé en WFS.*

Cependant, de par sa configuration impliquant un nombre considérable de haut-parleurs, la WFS n'est pour l'instant pas réellement implantée dans le paysage cinématographique actuelle [18] .

1.2.4 Tendances de mixage actuelles, et évolutions possibles

> *"La différence est que la façon dont un film va utiliser le surround peut maintenant être dictée par ce qui est approprié pour l'histoire plutôt que par la technologie. [...] N'importe quel son PEUT aller n'importe où dans la salle de cinéma, mais tous les sons ne DOIVENT pas aller n'importe où."*
>
> Mark Kerins [19] .

17. C'est une des raisons pour lesquels l'utilisation de la latéralisation et des surrounds et aujourd'hui assez limitée.

18. Pour de plus amples informations sur l'utilisation du système WFS au cinéma, vous pouvez vous reporter aux thèses de André (2013) et de Moulin (2015) ou au mémoire de fin d'étude de Remi Carreau et Thibault Macquart (2015).

19. À propos du passage aux canaux discrets et au son numérique au cinéma,*Beyond Dolby (stereo)*, Indiana University Press, 2011, p.71.

Aujourd'hui, l'esthétique générale des mixages en 5.1 reste relativement codifiée, et ce notamment dû aux habitudes liées au Dolby Stéréo et aux contraintes que le format imposait. On retrouve la plupart du temps le même genre d'organisation de l'espace sonore :

- Les dialogues sont placés quasi exclusivement dans l'enceinte centrale. Cela permet notamment à tous les spectateurs de percevoir la voix au même endroit (au centre de l'écran) et ce peu importe la place qu'il occupe dans la salle. L'effet ventriloque se charge d'associer les voix et les personnages si ceux-ci ne sont pas placés au centre de l'écran.

- Les bruitages et de nombreux effets sonores sont diffusés au centre avec les dialogues.

- Les ambiances nourrissent en général tous les canaux. La plupart du temps, les ambiances frontales et arrières sont décorrélées. On peut de plus noter que la tendance est de mettre des ambiance assez diffuses et légères dans les arrières, laissant les ambiances plus chargées et précises à l'avant.

- La musique est diffusée au minimum en stéréo avec régulièrement de la réverbération dans les arrières, ou alors la musique est directement mixée pour le 5.1.

- On retrouve parfois des effets dans les arrières ou se baladant dans l'espace surround [20], mais il s'agit essentiellement d'effets sonores "non réalistes" ou spectaculaires.

Si l'apparition des canaux discrets avec le son numérique a ouvert de réelles possibilités aux mixeurs, les nouveaux systèmes qui s'implantent actuellement proposent de nouvelles possibilités :

- Les trois systèmes (Atmos, Auro 3D et possiblement le WFS) proposent l'utilisation de la verticalité dans la bande son.

- On a pour la première fois une vraie égalité de niveau [21] et de bande passante entre les enceintes de surround et les enceintes frontales.

- La possibilité par le mixage objet de contrôler individuellement chaque enceinte et chaque objet sonore (pas dans l'Auro 3D).

- Ajout dans le Dolby Atmos d'enceintes surrounds proches de l'écran et de subwoofer à l'arrière de la salle.

Ainsi, ces nouveaux systèmes donnent un nouveau potentiel aux surrounds, mais augmentent également la possibilité de cohérence audiovisuelle spatiale, en azimut notamment. Le mixage objet semble permettre d'effectuer des placements et des mouvements de sources sonores beaucoup plus précis que pour les mixages classiques en 5.1. De la même manière que chaque nouveau format a toujours suscité des remises en question sur les possibilités du son au cinéma, ces nouveaux systèmes nous amènent à nous questionner sur l'utilisation

20. Surtout depuis le passage au son numérique et aux canaux discrets.
21. Alors que les enceintes surround étaient jusqu'alors calibrées à 82dB$_{spl}$(C), elles sont dans le Dolby Atmos étalonnées à 85dB$_{spl}$(C) comme les enceintes frontales.

de l'espace dans les bandes son actuelles. L'utilisation du mixage objet amène notamment la question de la cohérence audiovisuelle spatiale, et de l'apport éventuel des mouvements d'objets sonores, et plus particulièrement de la voix dans le cadre de notre mémoire.

1.3 La place de la voix au cinéma

Nous essaierons ici de voir la place que la voix a dans le cinéma. Nous serons autant intéressés par la place symbolique qu'elle tient dans la narration et dans la construction des films, que par la place physique qu'elle occupe dans l'espace sonore créé par le cinéma sonore.

1.3.1 Le cinéma qui parle

Si jusqu'en 1927 le cinéma était considéré comme "muet", il n'était pas pour autant sans dialogue, on qualifie même ce cinéma des débuts de très bavard. Si la voix ne se fait pas entendre, elle est déjà très présente. Et quand le cinéma devient enfin sonore, c'est avant tout pour donner une voix aux personnages ; et on parle souvent du cinéma "parlant". On reproche d'ailleurs souvent aux films des années 30 de n'être que de la parole filmée, et de ne voir le cinéma que comme une copie du réel sans véritable ambition artistique.

Mais si avec les bonimenteurs les répliques des comédiens étaient déjà parfois présentes dans les salles de cinéma, Mark Kerins dans son ouvrage *Beyond Dolby (stereo)* explique que c'est le couple que forment la voix et la bouche (et plus largement le corps) qui marque la vraie nouveauté dans le cinéma parlant. Une voix peut être alors associée au corps, et peut donc souligner un trait de caractère ou même volontairement s'opposer aux traits physiques d'un personnage. C'est ce que Dominique Sipière appellera le rapport de complémentarité, où le visage et la voix en harmonie s'enrichissent l'un l'autre, et le rapport d'opposition, où la voix s'oppose au corps (Il prend l'exemple de "Mr Bean, homme-enfant habité par la voix d'un géant") [22] . C'est cet écart parfois présent entre la voix et le corps qui a vu la carrière de certains acteurs du muet s'arrêter brutalement avec l'arrivée du parlant, leur voix ne correspondant pas aux attentes que les spectateurs et les producteurs avaient [23].

La parole qui fait sens

> *"L'interview est dépourvue de son, mais ce qui manque, c'est précisément la parole de la mère. C'est à dire le sens."*
>
> Jean-Louis Comolli [24] .

22. Dominique Sipière, "La voix au cinéma : divorces et retrouvailles", dans la revue *Tropisme* n˚17 "Questions de voix", 2011, p.166.
23. On pense évidemment au film *Chantons sous la pluie* de Stanley Donen (1952) basé sur le passage du muet au parlant et dont la star a une voix suraiguë n'allant pas du tout avec son corps.
24. "L'oral et l'oracle, séparation du corps et de la voix" dans la revue *Images documentaires*, "La voix", Premier semestre 2006, p.14.

En analysant, un passage du film *Intervista* de Anri Sala (1999), Comolli souligne le lien étroit qu'entretiennent au cinéma la voix et le sens. Le cinéma depuis qu'il sait faire entendre la voix, se construit autour de la parole. La voix et les mots qu'elle porte sont la charnière du cinéma tel que nous le connaissons, sa colonne vertébrale.

La Jetée de Chris Marker en est une très bonne illustration. Ce montage-photo est animée par une voix qui s'y superpose. C'est cette voix qui donne du sens aux images qui nous sont alors présentées, elle dirige l'image, l'oriente. Mais elle oriente également le regard du spectateur sur ces images.

Et même lorsque cette voix ne se fait pas comprendre elle n'en porte pas moins du sens. C'est le cas particulier de ce que Michel Chion appelle la "parole émanation". Présente essentiellement dans les films de Tati, la parole émanation est une parole incompréhensible dont l'intérêt n'est pas le propos, mais la voix en elle-même, la voix en tant qu'émanation d'un corps, d'un caractère. Ces voix qui ne disent rien, permettent à Tati de marquer les caractères de ses personnages, et de jouer sur les intonations plus que sur les mots en eux-même pour donner du sens.

De même, Michel Chion souligne que depuis sa création le cinéma est "vococentriste" : la voix et la parole prévalent sur le reste, et portent le sens de l'histoire. "*Il n'y a pas des sons, parmi lesquels entre autres, la voix humaine. Il y a la voix et tout le reste*." [25] Chion explique par là que c'est la voix qui hiérarchise les sons dans un film, et qui oriente notre écoute.

Si un certain type de cinéma spectaculaire très présent aux États-Unis échappe parfois à cette convention de cinéma vococentré, la plupart des films "narratifs", et notamment dans le cinéma français, se construisent autour de la voix. Aujourd'hui, l'absence de paroles devient une vraie intention artistique.

La voix sans corps

Là où l'arrivée de la voix au cinéma est une vraie avancée, c'est dans la possibilité de séparer la voix et le corps. Alain Boillat sépare les voix en trois catégories :

- La voix *in* qui vient d'un individu présent à l'image.

- La voix *off* dont la source est hors-champ, mais qui fait partie de l'univers diégétique.

- La voix *over* dont l'énonciateur est absent "en tant que source verbale" [26] de l'image et de la diégèse [27].

25. Michel Chion, *La voix au cinéma.*, Éditions de l'étoile/ Cahiers du cinéma (1982), réédition de 2005, p.18.
26. Alain Boillat, *Du Bonimenteur à la voix-over*, Antipodes, 2007, p.23.
27. Alain Boillat souligne qu'en France la voix-over est à tort appelée voix "off" par le grand public. Il préfère employer un terme différent pour éviter la confusion avec la traduction du mot anglais "off" qui signifie "hors-champ".

Comme le souligne Dominique Sipière, ces "lieux de la voix" communiquent entre eux. Aussi, la voix off nous donne souvent à voir quelques secondes ou minutes plus tard le corps correspondant. De même, une *voix-over* finit parfois par s'incarner sur un visage et devient alors une voix *in*[28].

La voix, souligne Jean-Louis Comolli, est ce qu'il reste quand on ne voit plus le corps, quand il est dérobé, c'est la trace sonore de ce corps. Désincarner la voix c'est permettre au cinéaste de jouer avec cette volonté qu'à l'homme de réunir la voix et le corps, de créer de la frustration.

C'est cette voix désincarnée qui prend parfois la forme de ce que Michel Chion nommera "l'acousmètre". L'acousmètre c'est une voix qui voit tout, qui entend tout, qui sait tout, qui est partout et qui se fait entendre de tous. Si Chion prend comme exemple la voix de la mère de Norman Bates dans *Psychose* (Alfred Hitchcock, 1960) ou la voix du parrain de la pègre du *Testament du docteur Mabuse* (Fritz Lang, 1933), Mark Kerins souligne la présence d'acousmètres dans des films plus récents. Ainsi, on peut penser à la voix de Frank dans *Donnie Darko* (Richard Kelly, 2001) ou encore à la voix du bouffon vert lorsqu'il parle à Norman Osborne dans une scène clé de *Spiderman* (Sam Raimi, 2002). Le bouffon vert est la deuxième personnalité de Norman Osborne qui prend peu à peu le contrôle de ce dernier. Dans cette scène, il lui parle à la fois dans sa tête mais aussi dans toute la pièce que Norman occupe, et c'est cette voix qui vient trouver enfin un corps avec le reflet de Norman dans son miroir.

FIGURE 1.14 – *La voix du bouffon vert s'incarnant dans le reflet de Norman Osborne (**Spiderman**, Sam Raimi, 2002).*

28. Sipière prend en exemple la voix over du narrateur, Anthony Hopkins, dans *Dracula* de Coppola qui vient ensuite correspondre au personnage de Van Helsing.

Souvent, la présence d'un acousmètre dans un film joue sur la volonté du spectateur (ou des personnages dans le cas de *Spiderman*) de découvrir qui se cache derrière l'acousmètre. Et si la voix trouve souvent un corps, il n'est pas rare que le visage reste masqué empêchant d'établir le lien entre la voix et la bouche, c'est le cas dans *Donnie Darko*. On peut également noter que quand la voix du bouffon vert se fait entendre en étant présent physiquement à l'image, c'est quand Norman porte son costume de super-vilain, il a donc le visage caché par un masque. Hormis l'artifice du miroir dans la scène d'introduction du méchant, il faut attendre la dernière réplique du bouffon vert pour que la voix du méchant trouve enfin une bouche. Dans cette scène, le bouffon vert enlève son casque et reprend la voix de Norman pour amadouer Spiderman. Mais juste avant de mourir, il reprend une ultime fois la voix sombre du bouffon vert, marquant définitivement que celui-ci a pris le pas sur la personnalité de Norman.

1.3.2 La voix spatialisée

> *"De la même manière que le silence ne peut exister que par rapport au son, mixer toutes les voix au centre a une signification différente dans un monde où elles pourraient être placées n'importe où."*
>
> Mark Kerins [29] .

L'évolution des technologies de reproduction au cinéma a ouvert de plus en plus la possibilité au mixeur de déplacer les sources sonores dans la salle de cinéma selon son désir. Si aujourd'hui, il est plus que courant de déplacer des effets sonores de manière très prononcée et de les envoyer dans les surrounds, la voix semble être dans la très grande majorité des cas cantonnée à l'enceinte centrale. Cependant, certains films se dérobent à cette convention.

Cohérence audiovisuelle spatiale en azimut

Si la voix devient temporellement synchrone avec l'image dès 1927, on ne peut que constater que spatialement le corps et la voix sont séparés. Si certains spectateurs furent d'abord troublés par cette séparation, la convention entra vite dans les habitudes. Dans ses premières expériences avec la stéréo dans les années 30, Blumlein réalise des petits films avec un couple de microphone placé sur la caméra, afin de recréer une scène sonore cohérente. Cependant, se pose le problème de la compatibilité de l'espace sonore avec l'espace visuel proposé par la caméra. Les valeurs de plan et l'ouverture du champ changent régulièrement, ce qui impose au champ sonore de varier en permanence. De plus, le montage déconstruit continuellement l'espace sonore créé par le couple.

L'arrivée des pistes magnétiques en 1950 et des grandes salles avec jusqu'à 5 enceintes derrière l'écran, entraine de nombreux essais de spatialisation des objets sonores : on essaie

29. En comparaison aux systèmes avec matriçage où la voix DOIT être au centre,*Beyond Dolby (stereo)*, Indiana University Press, 2011, p.258.

de réunir de manière cohérente la voix et son corps, mais les spectateurs, éduqués par vingts années de monophonie, se lassent vite de ces effets, les voyant comme une augmentation non pas de l'aspect réaliste d'un film mais de son aspect artificiel. L'arrivée du matriçage Dolby Stéréo obligera une fois pour toutes les ingénieurs du son à garder la voix dans l'enceinte centrale. Mais l'arrivée du son numérique remet en cause cet état de fait.

Le cas que l'on retrouve le plus dans le cinéma moderne est celui de la voix hors cadre, où la voix est alors spatialisée soit dans une des enceintes frontales Gauche ou Droite, soit parfois même dans les surrounds. Cette façon de mixer la voix sur les côtés permet au cinéaste de s'en servir comme un guide pour le spectateur, signifier la position du personnage, mais également le séparer de ce qui est IN. On peut penser au cas où une voix hors champ vient perturber une conversation se passant à l'écran, elle est alors parfois diffusée sur un côté de l'espace sonore. On peut noter que la voix revient la plupart du temps au centre une fois que l'on a découvert le personnage qui en était la source.

Si on a assez régulièrement de la cohérence audiovisuelle spatiale pour des sources hors cadre, c'est beaucoup plus rare pour des sources IN et donc frontales. Mais quelques films ont pris le parti d'avoir une cohérence entre le placement des voix et de l'image du corps.

Le mixeur de *Cars* (Pixar, 2006), Tom Myers, a durant tout le film fait correspondre la voix des personnages avec leur position à l'image. Dans la catégorie film d'animation, on retrouve également plus récemment le film de Remi Chayé mixé par Florent Lavallée, *Tout en haut du monde* (2015). Ce film présente également une cohérence scrupuleuse de la position des personnages et de leur voix. Mais cette cohérence en azimut s'accompagne systématiquement d'un respect absolu de la cohérence en profondeur et un travail impressionnant sur l'acoustique des lieux. Florent Lavallée dit cependant à propos de ce mixage avoir limité son utilisation des positions extrêmes gauche et droite pour éviter des positionnements possiblement critiques pour le spectateur, préférant garder le son "dans" l'écran.

Pour les films en prise de vue réelle il existe également des cas de cohérence en azimut. En 1995, le film *Strange Days* de Kathryn Biglow mixé par Gary Rydstrom utilise énormément de spatialisation des voix, cette esthétique se prêtant très bien au film basé sur un procédé d'illusion de la réalité. Plus récemment les films *Gravity* d'Alfonso Cuaron (2013) et *Birdman* d'Alejandro Inarritu (2014) ont adopté cette esthétique.

Dans *Birdman* c'est véritablement les placements des voix et des sons qui dirigent le regard, mais surtout qui dirigent les mouvements des personnages et de la caméra. Il est cependant intéressant de se rendre compte que les spatialisations des voix se font quasi exclusivement quand celles-ci sont hors champ ou quand elles sont au bord du cadre, avec souvent un jeu sur les entrées et les sorties de champ. Mais dès que le discours des personnages est

réellement important, ou si les personnages sont à des positions intermédiaires entre centre et côté, la voix revient systématiquement dans l'enceinte centrale.

Si ces films ont tous exploité avec succès la cohérence audiovisuelle spatiale en azimut, il est cependant important de souligner que ce sont tous des cas particuliers. Deux sont des films d'animations. De plus, pour *Cars* Cormac Donnelly souligne que les sons de voitures sont souvent spatialisés de manière cohérente, et qu'il était donc logique de faire la même chose pour les voix associées [30]. *Strange days*, *Gravity* et *Birdman* sont également des cas à part. Les deux derniers sont basés sur l'utilisation de plans séquences et ne sont pas bousculés par un découpage continuel de l'action. Et le premier se base sur un procédé incitant à la cohérence en azimut, puisque l'on a de nombreuses séquences en "point of view" [31]. La cohérence permet d'ailleurs à Gary Rydstrom de séparer les moments en "point of view" des moments de narration "classiques" mixés au centre. On ne peut donc pas faire de généralité sur l'intérêt de la cohérence audiovisuelle spatiale pour la voix.

Le cas particulier : Les voix sans corps

Dans le cinéma d'aujourd'hui la spatialisation des voix est très récurrente quand ces voix ne sont pas présentes physiquement à l'écran. La discrétisation des canaux, notamment surround, donne une solution pour séparer ces voix qui n'ont pas le même statut que le reste des dialogues du film. Aussi, très souvent les voix dites "intérieures" sont spatialisées dans les surrounds ou dans la salle de cinéma, afin de les placer dans un lieu qui n'est pas complètement celui de la réalité. On peut prendre comme exemple les voix qui hantent le personnage de Max dans *Mad Max : Fury Road* de George Miller (2015), les voix se baladent tout autour de la salle pour faire ressentir le trouble psychologique du personnage.

Les surrounds permettent également de donner une nouvelle dimension aux acousmètres. Ainsi, ce dernier peut vraiment "être partout". Par exemple, dans la scène où Norman entend pour la première fois la voix du bouffon vert, celle-ci raisonne dans toute la pièce et change instantanément d'emplacement. Le montage image qui montre Norman chercher la provenance de la voix renforce cette maîtrise de l'acousmètre qui peut changer de place à sa guise. Et comme Norman, le spectateur ne sait pas où se trouve cette voix, soulignant ainsi la folie grandissante chez Norman.

On peut penser également à la voix du personnage imaginaire "Birdman" qui pendant tout le film s'adresse à Michael Keaton dans sa tête. Cette voix, qui n'est ni hors de l'histoire, ni réellement présente, est mixée dans tous les canaux, donnant une sensation d'englobement totale. Ce n'est que dans un des délires du personnage où Birdman apparaitra à l'ecran, que sa voix sera alors mixée au centre.

30. "Dialogue on the move-panning in Gravity, Cars and Strange Days" article sur Designingsound.org.
31. Illusion que le cadre de la caméra correspond à la vision d'un personnage.

FIGURE 1.15 – *La voix de Birdman présente jusqu'alors uniquement dans la tête de Michael Keaton se matérialise à l'écran (**Birdman**, Alejandro Inarritu, 2014).*

Quelles limites ?

Si plusieurs films ont tenté la cohérence audiovisuelle spatiale, la majorité des films gardent la voix au centre. On peut trouver plusieurs explications à cette tendance :

- Garder la voix au centre permet que tous les spectateurs perçoivent la voix au même endroit. En effet, si le mixeur place par exemple la voix entre l'enceinte centrale et l'enceinte de gauche, un spectateur situé à gauche de la salle percevra la voix sur l'enceinte de gauche et non entre les deux enceintes.

- L'effet ventriloque, qui correspond à la capacité du cerveau de lier un événement sonore et un évenement visuel spatialement séparés, permet au spectateur de ne pas être gêné par la non cohérence de positionnement de la voix.

- Dans le 5.1 (et dans le 7.1 à moindre échelle), les canaux arrières sont diffusés sur un ensemble de haut-parleurs répartis dans la salle, ce qui empêche d'avoir un positionnement précis dans les surrounds. De plus, les arrières commencent à 1/3 de la salle, il est donc difficile de faire des suivis corrects entre l'avant et l'arrière de l'espace sonore.

- Historiquement, l'enceinte centrale est l'enceinte qui était la plus utilisée et donc la mieux entretenue par les exploitants. Le risque de déplacer la voix, c'est que les enceintes ne soient pas réglées de la même manière. Les déplacements de voix peuvent être soulignés par des changements de timbres non souhaités. De plus, il arrive que certains cinémas ne remplacent pas nécessairement les enceintes en panne quand ce n'est pas l'enceinte centrale. Déplacer la voix, c'est donc prendre le risque que certaines paroles ne soient pas diffusées comme il faut

dans la salle. Vincent Arnardi m'a rapporté une mésaventure similaire durant la projection d'un film où il avait fait suivre la voix d'un personnage avec son déplacement à l'écran. Les enceintes latérales étant hors services, la voix avait alors disparu le temps de quelques secondes.

- Le montage ne cesse de malmener la continuité temporelle des séquences, et c'est généralement le son, et par association la voix, qui permet d'assurer une cohérence temporelle. La cohérence audiovisuelle spatiale sur des films avec des montages classiques briserait cette continuité amenée par le son, et soulignerait l'artifice du découpage.

- Le cinéma n'a pas pour but de recréer le réel, mais est un moyen d'expression à part entière. La cohérence audiovisuelle spatiale permettant de se rapprocher d'une "réalité potentielle" n'est donc pas nécessairement appropriée pour le cinéma en tant qu'art.

Cependant, nous avons vu dans la partie 1.2.4 que les nouveaux systèmes de diffusion proposent des moyens diminuant les contraintes quant à la spatialisation de la voix. Ainsi, les salles sont de mieux en mieux équipées et bien entretenues (c'est déjà le cas dans les salles modernes), ce qui tend à réduire les problèmes de réglages des salles.

De plus, l'arrivée du mixage objet permettrait de faire des mouvements plus fluides et de pouvoir diffuser le son directement sur une enceinte des surrounds et non plus sur un ensemble de haut-parleurs. Le retour des enceintes Inter-Gauche et droite (Dolby Atmos et 22.2) pourraient diminuer l'impact de la position du spectateur dans la salle, et la solution de la WFS proposerait une vrai possibilité de s'extraire de ce problème, puisqu'elle propose une zone d'écoute élargie.

Malgré tout, il reste un problème majeur : "L'Exit Sign Effect". Cela correspond à l'effet produit par un son trop spatialisé qui sortirait le spectateur de l'illusion cinématographique, par exemple en se retournant (à cause d'un son dans les arrières) et en apercevant le panneau "exit" de la salle de cinéma. Or, nous avons souligné que dans le cinéma actuel, c'est la voix qui est la colonne vertébrale de la bande son. On peut donc supposer que les effets qui déplacent la voix sont encore plus susceptibles de provoquer cet "exit sign effect". Et cela s'étend également à des effets frontaux, avec l'écueil du déplacement de son sur un champ contre-champ (montré notamment par l'exemple de Blumlein).

Il apparaît donc intéressant de creuser la question de la cohérence audiovisuelle spatiale de la voix. Peut-on vérifier que les spectateurs ne jugent pas de la même manière la voix et les autres objets sonores ? Le montage image est-il rédhibitoire pour l'utilisation de la cohérence audiovisuelle en azimut ? On peut aussi imaginer que, comme par le passé, le spectateur va s'habituer à une utilisation marquée de la cohérence spatiale en azimut. Ce sont toutes ces interrogations qui animeront le test perceptif réalisé dans le cadre de ce mémoire.

Chapitre 2

Etat de l'art : perception image et son

Avant d'aller plus loin sur la place de la voix au cinéma, et d'entamer mon test perceptif, il apparaît important de faire un rappel sur les relations qu'entretiennent la perception auditive et la perception visuelle. Nous aborderons d'abord l'impact de l'image sur la perception du son par l'étude des biais intersensoriels et de l'effet ventriloque. Ensuite, nous traiterons des différents tests perceptifs ayant abordé l'impact de la cohérence du son à l'image.

2.1 L'effet Ventriloque

2.1.1 Biais Intersensoriels

Dans notre vie de tous les jours, nous analysons constamment des événements multimodaux, c'est-à-dire qui font entrer en jeu différentes modalités de perception. Il apparait intéressant de s'interroger sur la manière dont l'homme analyse ces évènements mais également de voir le poids que possède telle ou telle modalité sur cette analyse ; le son et la vision possèdent-ils le même poids dans le compréhension d'un événement audio-visuel ?

De nombreux chercheurs ont tenté de comprendre les mécanismes qui régissent les interactions entre ces différentes modalités de perception. On dit souvent que la vision domine les autres sens, le but est alors de comprendre les influences réelles qu'ont les modalités perceptives les unes sur les autres. Afin de saisir ces influences, les scientifiques utilisent ce qu'on appelle parfois des situations de conflits sensorielles. Les sujets des tests en question se voient présenter deux informations contradictoires selon la modalité perceptive, autrement dit on crée des disparités entre une perception et une autre.

La plupart du temps, nous sommes confrontés à des évènements qui font appel à plusieurs modalités donnant la même information de localisation. Mais il arrive parfois que les informations de localisation apportées par les différentes modalités d'un événement ne soient pas identiques, c'est le cas notamment dans le cadre d'une projection cinématographique où

la position de l'enceinte qui délivre le son ne correspond pas nécessairement à la position du personnage qui parle à l'écran. Ce sont ces situations que tentent de comprendre ces études.

Dans les études visant à mesurer les biais intersensoriels, les sujets sont soumis à des tâches dites de localisation. Cela signifie qu'ils seront amenés à localiser un stimulus visuel ou auditif.

Biais de l'audition par la vision

Dans leur étude, Pick et al (1969) crée grâce à un prisme optique un déplacement de la vue d'un haut-parleur de 11˚ latéralement, créant ainsi une disparité par rapport au son de ce haut-parleur. Les résultats montrent un biais de l'audition par la vision égal à 48 % . Ce pourcentage représente un biais relatif, c'est-à-dire le rapport entre la séparation spatiale des deux stimulus (ici 11˚) et le biais absolu, autrement dit l'angle moyen de l'erreur de localisation du stimulus cible [1] par les sujets.

Pick et al n'ayant fait la mesure que pour une angulation d'écart, Bertelson et Radeau (1981) ont fait des mesures de biais pour des disparités de 7˚, 15˚ et 25˚. Les biais intersensoriels absolus obtenus par les sujets étaient respectivement de 4˚, 6.3˚ et 8.2˚ (relatifs : 57, 41 et 32 %), autrement dit les stimuli sonore ont été localisés décalés vers le stimulus visuel.

Thurlow et Weerts (1971) ont montré une influence de la direction des yeux, et de l'attente des sujets quant au placement de la source sonore. L'expérience consistait en la diffusion d'un son en face des sujets par un haut-parleur dissimulé, et un haut-parleur visible 20 degré sur le côté. Les sujets regardant le haut-parleur visible et à qui on avait expliqué que le son viendrait de ce haut-parleur présentaient un biais de 9˚, ce qui était plus important que lorsqu'on ne leur annonçait pas que le son viendrait du haut-parleur visible. De même, quand les sujets regardaient devant eux (donc vers le haut-parleur caché), on n'observait pas de biais vers la position du haut-parleur visible, bien qu'on leur ait dit que le son viendrait de ce haut-parleur.

Il existe deux théories concernant la manière dont les humains localisent des évènements multimodaux contradictoires. La première est la théorie de la modalité pertinente (aussi appelée parfois le phénomène de capture visuelle). Cette théorie consiste à dire que la modalité la plus précise dominera complètement le jugement de localisation de l'événement multimodal. La vue, étant beaucoup plus précise dans sa définition spatiale que le son, sera donc dominante (d'où le nom de capture visuelle).

L'autre théorie est celle de l'estimation du maximum de vraisemblance (Maximum Likelihood Estimation ou MLE en anglais). Cette théorie prend en compte la précision de localisation

1. Pour mesurer le biais de l'audition par la vision, le stimulus cible sera le stimulus sonore.

de chaque modalité de manière unimodale. L'influence de la localisation selon chaque modalité (la vue et l'ouïe) sur la localisation de l'évènement audiovisuel sera pondérée en fonction de leur pertinence, autrement dit de leur précision en condition unimodale.

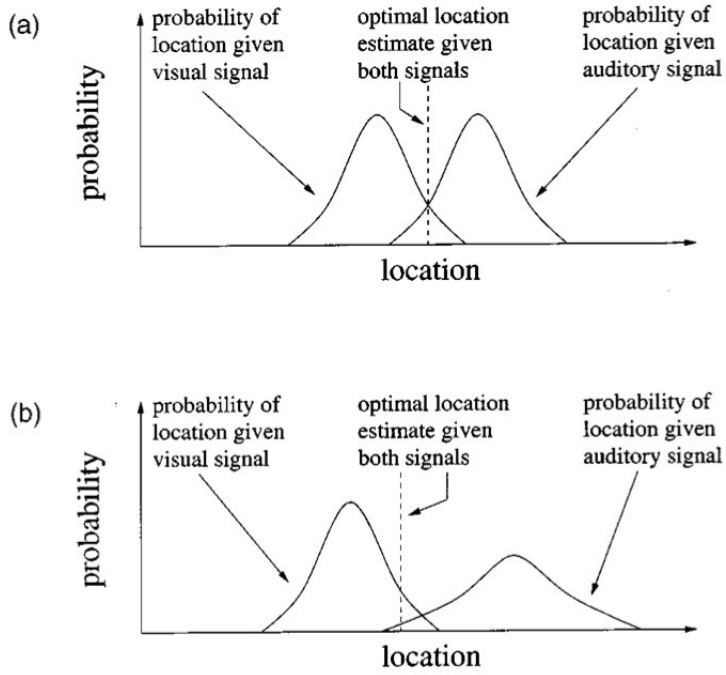


FIGURE 2.1 – *Représentation de la théorie du maximum de vraisemblance pour un événement audio-visuel. (a) Représente un cas où la pertinence de chaque modalité est équivalente, (b) une situation où la pertinence du stimuli visuel est plus importante (Battaglia 2003)*

La figure 2.1 représente en abscisses la localisation, et en ordonnées la probabilité que l'événement se produise à une localisation donnée. On peut se rendre compte que les deux modèles ne sont en fait pas incompatibles. Comme le souligne Battaglia (2003), la théorie de la modalité pertinente représente un cas particulier de la théorie du maximum de vraisemblance, où le poids d'une modalité serait extrêmement supérieur au poids de la modalité en conflit.

Selon le modèle du maximum de vraisemblance, la localisation L d'un événement s'obtient de la manière suivante :

$$L = w_v L_v + w_a L_a \tag{2.1}$$

avec

$$w_v = \frac{\sigma_a^2}{\sigma_v^2 + \sigma_a^2} \text{ et } w_a = \frac{\sigma_v^2}{\sigma_a^2 + \sigma_v^2} \tag{2.2}$$

Ici, L correspond à la localisation de l'évènement multimodal, L_v et L_a correspondent à la localisation de l'évènement selon la modalité visuelle ou auditive, σ_v^2 et σ_a^2 représentent la variance de la localisation selon la modalité visuelle ou auditive, enfin w_a et w_v représentent la pondération de la modalité sur l'événement multimodal.

Ce modèle laisse envisager la possibilité que l'audition puisse avoir une influence sur la perception de la vision, si l'audition est plus précise que la vision. Il apparaît effectivement dans différentes études que l'audition peut dans certains cas biaiser la vision. Dans l'étude de Pick et al (1969) on obtient un biais de la vision par l'audition de seulement 1% ce qui laisse indiquer que la vision dominerait totalement l'audition. Cependant, d'autre études (Alais et Burr 2004, Warren et Welch 1981) ont montré une influence significative de l'audition sur la vision. En 1981, Warren et Welch, dans une étude portant sur l'impact du caractère incontestable du lien entre les stimuli auditifs et visuels discordants sur l'effet ventriloque, ont montré un biais de la vision par l'audition de 17 % en moyenne, avec une séparation de 10˚ entre les stimuli visuel et sonore.

Dans son étude sur la véracité des théories du maximum de vraisemblance et de la modalité dominante, Battaglia (2003) a montré qu'en réduisant la précision de l'information visuelle (en y ajoutant plus ou moins de bruit) le jugement de localisation se faisait de plus en plus selon la perception auditive. Au plus haut niveau de bruit testé, le jugement de localisation se faisait à égalité par les informations spatiales apportées par le son et par la vision. [2]

Alais et Burr en 2004 montrent également que la précision de l'information visuelle quant à la localisation de l'événement bimodal est primordiale pour qu'il s'établisse ou non une capture visuelle. En effet, ils ont effectué des tâches de localisation dans des situations de conflit entre la vision et l'audition, en faisant varier la netteté et la largeur de l'information visuelle. Les stimuli étaient des taches lumineuses (plus ou moins floues et larges) associées à des clics sonores de 1,5ms. Pour une tache de largeur de 4˚, l'information visuelle dominait totalement la localisation bimodale. Mais avec une largeur autour de 30˚ les deux modalités avaient une importance équivalente sur la perception de l'événement bimodal. Enfin, avec une largeur de la tâche autour de 60˚, c'est l'audition qui devenait plus précise et qui donc dominait la décision. C'est ce que les auteurs appelèrent l'effet ventriloque inversé ; cet effet ayant été montré pour un écart entre les stimuli sonores et visuels de 10˚ dans cette étude.

2. Il a également montré que si le modèle MLE avait plutôt bien prédit la courbe de réponse en fonction de l'augmentation du bruit dans l'information visuelle, il ne prenait pas assez en compte l'importance de la vision. C'est pourquoi l'étude propose l'utilisation d'un modèle Bayésien qui prévoit mieux l'évolution des biais intersensoriels.

2.1.2 L'effet ventriloque en azimut

L'effet ventriloque est une des raisons parfois avancées par les mixeurs pour expliquer que les dialogues, ainsi que beaucoup d'objets sonores, sont souvent mixés dans l'enceinte centrale. L'effet ventriloque correspond à la capacité du cerveau à corriger les disparités spatiales d'un stimulus sonore et d'un stimulus visuel associé. C'est le cas classique du ventriloque qui crée l'illusion que la voix vient de la marionnette, en agitant les lèvres de celle-ci.

Pour étudier le phénomène de fusion entre un stimulus visuel et un stimulus sonore séparés dans l'espace, la tâche de localisation n'est pas forcément appropriée. Radeau et Bertelson (1981) ou encore Thrulow et Jack (1973) recommandent plutôt des tâches dites de discrimination. Le but étant de diminuer la trop grande différenciation qu'amène une tâche de localisation. Ici, les sujets ne devront pas indiquer la direction du stimulus, mais devront déterminer s'ils considèrent que les deux stimuli sont émis ou non du même endroit, ou s'ils "fusionnent".

Certains auteurs ont plutôt utilisé l'échelle de dégradation ITU-R à 5 notes, permettant de juger à quel point les disparités sont perceptibles et gênantes. Chacune des 5 notes est associée à une sensation quant à la gêne ressentie ou non par le sujet :
- 1 = Très gênant
- 2 = Gênant
- 3 = Légèrement gênant
- 4 = Perceptible mais non gênant
- 5 = Imperceptible

Afin de déterminer la limite angulaire acceptable, Komiyama a par exemple situé la limite entre "perceptible mais non gênant" et "légèrement gênant" (Komiyama 1989).

Ce sont Howard et Templeton qui les premiers parlent d'un effet dit "ventriloque" en 1965. Depuis, de nombreuses études ont testé différents critères influant ou non sur la présence de cet effet de fusion entre deux stimuli séparés dans l'espace. La littérature sur le sujet porte néanmoins à confusion : en effet, le terme d'effet ventriloque a très souvent été utilisé pour des études portant sur des mesures de biais intersensoriels et non une réelle fusion entre les stimuli visuels et auditifs.

Warren et Welch (1981) suggèrent que l'effet ventriloque ne serait qu'un cas particulier des biais intersensoriels, où la somme du biais de l'audition par la vision et inversement serait de 100 %. Aussi, ils le montrent sur un dispositif présentant 10˚ de séparation entre les stimuli visuels et auditifs. Ils obtiennent un biais de l'audition par la vision de 79%, et un biais de la vision par l'audition de 17%, la somme valant alors 96%, signifiant un effet de fusion. On peut donc y voir une façon de lier l'effet ventriloque et les biais intersensoriels.

Toutes les études observent une diminution de l'effet avec l'augmentation de l'écart angulaire entre les stimuli sonores et visuels. Cependant, les résultats quant à la séparation maximale pour laquelle on rapporte une fusion diffèrent selon les études. En effet, les résultats s'étendent de 3˚ pour Lewald en 2001 jusqu'à 30˚pour Jackson dans une des premières études sur le sujet en 1953. La différence entre les résultats s'explique par la différence entre les expériences, que ce soit le dispositif, les stimuli utilisés, la question posée aux sujets, ou encore le réalisme de l'association des stimuli visuels et auditifs.

Vraisemblance de l'association image et son

Le critère qui me semble le plus intéressant à relever, et qui a d'ailleurs fait l'objet de plusieurs études, est le réalisme de l'association entre les stimuli visuels et sonores. Les études s'accordent à dire que plus la combinaison entre les stimuli sonores et visuels était réaliste et vraisemblable, plus l'effet ventriloque était important, et inversement (Jackson 1953, Thurlow et Jack 1973, Warren et Welch 1981, Radeau et Bertelson 1977). Cela parait effectivement logique, puisque plus le stimulus sonore paraitra correspondre au son que pourrait produire le stimulus visuel, plus il sera aisé pour le sujet de considérer que les deux stimuli constituent un seul stimulus audiovisuel, et donc l'effet de fusion spatiale sera facilité.

Dès 1953, dans l'étude de Jackson, c'est un des critères étudiés. Ce dernier teste l'effet de fusion sur deux associations différentes :

- Dans un premier temps le stimulus visuel était de la vapeur d'eau s'échappant d'une bouilloire, et le son correspondant était un sifflement typique de bouilloire.

- Le deuxième stimulus associait un son de cloche à des allumages de lumière.

Les deux associations n'ont pas obtenu les mêmes résultats du tout. Alors que le son de sifflement était perçu "sur" la bouilloire dans 97% des cas jusqu'à 30˚d'écart ; pour l'association de lumière et de son de cloche, dès 22.5˚de disparité, la lumière et le son de cloche ne fusionnait que dans 43% des cas.

De même, dans l'étude de Thurlow et Jack (1973), un dispositif était présenté mettant en place un écran vidéo et une enceinte située 20 ˚à gauche de la télé. Le stimulus sonore, d'une durée de 5 minutes, était une voix d'homme et le stimulus visuel associé était soit une bouche synchrone avec le texte lu, soit une croix blanche au centre de l'écran. Les sujets devaient indiquer en appuyant sur un interrupteur dès qu'il percevait le son comme venant du stimulus visuel :

- Pour l'association voix-bouche, tous les sujets ont perçu le son comme venant de la bouche, avec une durée moyenne de 3 minutes et 22 secondes (/5min).

- Pour l'association avec la croix blanche, seulement 2 sujets sur dix ont déclaré avoir senti le son venir de la croix, avec une durée moyenne de 51 secondes (/5min).

Dans une autre étude, Thurlow et Jack effectuait un test similaire avec la diffusion d'une image de marionnette comportant plus ou moins de détails faciaux (yeux, bouches, nez) et avec présence ou non de mouvements de la marionnette. Ils ont montré que plus le nombre d'indices visuels étaient important, plus la fusion durait pour un temps important, de même le mouvement de la marionnette semblait être encore plus pertinent que la présence d'indice visuel d'un visage.

Warren et Welch en 1981 ont testé pour une disparité de 10˚entre le stimulus visuel et le stimulus auditif différents niveaux de vraisemblance. Une voix était diffusée avec une image de visage ou avec uniquement la télé allumée sans image. Ici, les sujets devaient effectuer une tâche de localisation. Pour le premier, cas le biais total (auditif + visuel) était de 96.8% , on peut donc considérer qu'il y a fusion. A l'inverse pour la deuxième condition le biais n'était que de 31.9%.

Il semblerait donc que l'effet ventriloque soit extrêmement robuste lorsque l'association entre les stimuli visuels et auditifs est "naturelle" ou du moins vraisemblable. C'est le cas que l'on rencontre dans la très grande majorité dans la vie de tous les jours, ou bien dans le cas d'une projection cinématographique.

Différence de temps entre stimuli sonore et visuel

Plusieurs études ont montré que retarder le stimuli sonore par rapport à un stimulus visuel placé à un endroit différent, diminuait l'effet ventriloque (Thurlow et Jack 1973, Warren et Welch 1981, Wallace 2004).

Ainsi, Thurlow et Jack (1973) toujours avec le même dispositif (une télévision droit devant le sujet et une voix diffusée 20 ˚sur la gauche) ont montré qu'en augmentant le retard du son par rapport à l'image, l'effet de fusion durait de moins en moins longtemps. Avec un retard de 100ms, les sujet rapporte une fusion pour 220 secondes (sur 300 seconde possible au maximum) [3]. Pour un retard de 200ms il n'y a plus que 114 secondes de fusion, et pour un retard de 300ms uniquement 21.3 secondes de fusion.

Dans son étude, Wallace (2004) teste l'effet ventriloque pour 3 retard différents (200, 500 et 800ms) et 3 disparités entre stimuli visuel et auditif (5˚, 10˚, et 15˚), avec des flashs lumineux, associés à des salves de bruit. Pour un retard de 200ms, même à 15˚, plus de 54% des sujets rapportent une unité entre les deux stimuli. Alors que pour 500ms et 800ms, nous n'avons plus que respectivement 41.5% et 37.5 % des sujets qui rapportent une fusion, le seuil à 50% étant alors ramené à 10˚.

3. Les sujets appuient sur un interrupteur tant qu'ils considèrent que les deux stimuli fusionnent.

De même, Warren et Welch, en retardant le son de 150ms, passent d'un biais total de 96.8% à seulement 40.2%, pour une séparation de seulement 10˚.

Différence de dispositif/protocole

Différentes études ont montré que les résultats étaient très dépendant de la tâche demandée aux sujets. Par exemple, dans l'expérience de Warren et Welch (1981) on obtenait un biais total presque égal à 100% (voir plus haut). Sur ces résultats, les sujets étaient informés, à tort, qu'il n'y avait qu'un seul événement audiovisuel. Lorsque l'on indiquait aux sujets qu'ils étaient en présence de deux stimuli séparés, le biais tombait à 35.5%.

De même, plus la tâche demandée au sujet est précise, moins l'effet ventriloque sera important. C'est pourquoi Bertelson et Radeau déconseillent les tâches de localisation. Aussi les résultats très faibles de Lewald (un effet ventriloque de seulement 3˚) s'expliquent notamment par son protocole très compliqué et demandant une extrême concentration aux sujets. Ces derniers voyaient un laser se déplacer et le son était diffusé par une enceinte cachée. Les sujets devaient appuyer sur un bouton quand ils considéraient que le laser arrivait à l'endroit d'où le son provenait.

Impact de l'expérience des sujets

Il semblerait que l'expérience du sujet ait un impact très important sur la robustesse de l'effet ventriloque. En effet, dans l'étude de Komiyama (1989), la limite acceptable de séparation entre les stimulus visuel et sonore pour des sujets naïfs [4] était de 20˚, alors que pour des sujets experts [5] le seuil descendait à seulement 11˚.

Cela souligne bien que l'écoute d'un expert diffère énormément d'une écoute naïve, puisque l'expert sera beaucoup plus analytique dans son approche, alors qu'un sujet naïf sera plus spontané et naturel.

2.1.3 L'effet ventriloque en élévation

Nous présenterons ici succinctement les différents résultats obtenus sur les rares études portant sur l'effet ventriloque en élévation.

En 1973 dans leur étude sur l'effet ventriloque, Thurlow et Jack ont testé celui-ci sur le plan vertical. Ceux-ci ont observé des fusions pour des angles de 55˚et 195˚(l'enceinte étaient placée derrière les sujets), les durées de fusion [6] obtenues étaient respectivement de 246sec

4. On parle de sujets "naïfs" quand ceux-ci ne sont pas en lien avec le champ d'action du test. À opposer aux sujets dits "experts"..

5. Ici des ingénieurs du son et des acousticiens

6. Comme pour l'effet ventriloque en horizontal, les sujets devaient appuyer sur un bouton lorsqu'ils sentaient une fusion entre la voix et l'image de l'homme.

et 199sec (pour un stimulus d'une durée de 5 minutes). Cette étude semble donc indiquer une extrême robustesse de l'effet ventriloque en élévation.

Dans l'expérience IV de sa thèse, Etienne Hendrickx a également présenté l'image d'un homme déclamant un texte au centre d'un écran de projection. La projection était effectuée en 3D. Les limites de fusion en vertical s'étalaient pour les sujets de 23˚pour les plus discriminant et jusqu'à 137˚pour certains. Or, Hendrickx montre que dans une application au cinéma, même un sujet au premier rang aura une ouverture verticale de son champ de vision inférieure à 20˚(il l'estime à 19.4˚). Ces différents résultats impliquent que la cohérence audiovisuelle spatiale en élévation au cinéma n'est peut-être pas pertinente, tant l'effet ventriloque semble déjà extrêmement robuste dans cette dimension.

2.1.4 Discussion

Nous avons montré que le cerveau humain était capable de corriger des disparités assez importantes de placement entre deux stimuli associés.

Il semble néanmoins intéressant de souligner que l'ensemble de ces expériences sont effectuées dans des conditions très éloignées de la réalité de l'exploitation audiovisuelle. Tout d'abord, les tâches demandées aux sujets rendent le test forcément non neutre par rapport au paramètre étudié. De plus, on se rend compte que moins les tâches demandées aux sujets sont précises, plus l'effet ventriloque semble important.

De même, les conditions physiques de test sont très éloignées d'une consommation classique d'un produit audiovisuel. Les sujets sont la plupart du temps installés dans des sièges de dentiste, ou des sièges empêchant de bouger. De plus, les mouvements de la tête sont la plupart du temps maitrisés par l'expérimentateur, par divers moyens.

Il est donc fort à parier que l'effet ventriloque est encore plus fort dans des conditions normales de diffusion d'un stimulus audiovisuel, que ce soit au cinéma ou dans d'autre cas.

2.2 Cohérence audiovisuelle spatiale des éléments sonores

Le but de cette partie est de présenter les différentes expériences ayant déjà été mise en place et mettant en jeu des situations plus réalistes et moins abstraites que les études sur l'effet ventriloque. Il n'existe pas beaucoup d'expériences ayant réellement abordé la cohérence de placement des objets sonores et notamment de la voix, il s'agira ici de présenter leurs résultats. Nous aborderons tout particulièrement une des expériences effectuées par Etienne Hendrickx, qui sera la base de mon propre test perceptif.

2.2.1 Différentes études sur la cohérence audiovisuelle spatiale

Dans cette première sous-partie, nous parlerons des différentes expériences ayant abordé l'impact de la cohérence audiovisuelle spatiale. Si la cohérence azimutale n'est pas nécessairement le centre de leurs études, les résultats soulignent certains points intéressants, tant sur les résultats à proprement parler, que sur les protocoles employés.

Cohérence audiovisuelle en distance

Plusieurs études ont abordé l'influence de la cohérence audiovisuelle en distance, notamment pour le cinéma stéréoscopique (3D), nous présenterons donc rapidement les résultats de ces expériences.

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

Dans sa thèse, portant sur le son spatialisé pour la vidéo 3D, ainsi que l'intégration de la WFS dans une expérience audiovisuelle 3D, Samuel Moulin effectue une expérience portant sur l'impact du rendu de la distance sonore sur l'expérience du spectateur. Il utilise pour cela neuf séquences en 3D. Moulin effectue trois mixages différents :
- Un mixage dit "sans distance" : tous les éléments sonores sont restitués sur le même plan, mais en respectant l'azimut des objets visuels correspondants aux sons.
- Un mixage "distance réaliste" qui synthétise les objets sonores à la positon des objets visuels (azimut et distance) avec des positions y compris devant l'écran.
- Un mixage "distance augmentée" où on a le même azimut mais des informations de distance amplifiées.

Les résultats du test ont permis d'observer que :
- Sur l'évaluation d'un critère de profondeur, les notes ont été globalement plus élevées pour les mixages avec distance réaliste ou augmentée.
- Les différents mixages n'ont pas influencé les critères de qualité sonore et de gêne visuelle.
- Enfin, uniquement deux séquences sur neuf ont présenté des notes positives quant à la sensation d'immersion apportée par la cohérence en profondeur.

Cette expérience semble donc souligner la très faible influence de la cohérence audiovisuelle en profondeur sur l'expérience audiovisuelle globale des spectateurs.

Cependant, on peut remarquer que les séquences employées durant le test étaient très peu variées. Toutes présentent deux personnages, dans un seul lieu, et avec un seul et unique cadre. De même, les séquences étaient également cohérentes également en azimut. Or, ce critère n'est pas pris en compte par le test, qui n'étudie que la profondeur. On ne peut pas savoir

si l'absence de différence entre les notes des séquences ne peut pas provenir par exemple du trouble des sujets face à une spatialisation cohérente des voix en azimut.

- Kruszielski et al (2012) : *Évaluation 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.*

Dans cette étude, l'expérimentateur présentait un saxophoniste jouant dans un studio à 5 distances différentes d'une caméra 3D.

Le son correspondant n'a pas été enregistré de manière synchrone, mais en prise de son de proximité dans un autre studio moins réverbérant. L'acoustique du premier studio a ensuite été simulée, et une technique utilisant les microphones virtuels a permis d'obtenir 5 stimuli sonores plaçant le saxophoniste aux 5 positions enregistrées au préalable par la caméra.

Les différents stimuli sonores étaient mélangés aux stimuli visuels. Les sujets devaient juger à quel point le son de la séquence était adapté à l'image, sur une échelle allant de 1 "pas adapté du tout" à 7 "très adapté".

Les résultats montrent que la position sonore la plus adaptée pour chaque séquence était celle qui simulait la même distance au son qu'à l'image. Et inversement, plus le stimuli sonore était à une distance qui s'éloignait du saxophoniste à l'image, moins le son semblait adapté.

Si dans l'étude de Moulin, la cohérence en profondeur ne semble pas être un facteur déterminant, dans l'étude de Kruszielski les résultats semblent montrer que plus la position de l'objet sonore est cohérente plus le son semblera adapté. Cependant les différences entre ces études peuvent trouver une explication. En effet les stimuli utilisés par Moulin tentait de mettre en scène des petites séquences jouées par des comédiens, alors que dans l'étude de Kruzielski le stimulus se résume à un musicien placé à différentes distances. De plus, si dans l'étude de Moulin, on avait également une cohérence en azimut, ici le stimulus était placé au centre et il n'y avait que le paramètre de profondeur qui jouait.

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

Dans cette étude, André étudie l'impact de la cohérence audiovisuelle spatiale dans un contexte cinématographique sur la sensation de présence ressentie par le spectateur. Par sensation de présence on entend la sensation d'être "présent" dans le film, on pourrait parler d'immersion.

Les séquences utilisées dans ce test sont extraites d'un film d'animation en 3D, et non d'images en prise de vue réelle. Les séquences sont diffusées par un système WFS. Trois versions différentes sont présentées :
- Une version avec une faible cohérence spatiale (une version stéréo).
- Une version très cohérente spatialement, qui consiste en une version complètement orientée objet et rendue par le système WFS.
- Une version hybride, qui correspond à une optimisation d'une version stéréo avec les métadata de la version orientée objet. C'était donc une version stéréo avec cohérence azimutale et en profondeur optimisée.

Les résultats ont montré une absence d'impact de la version de mixage sur la sensation de présence. Cependant, André a isolé un groupe de 12 sujets (sur 33) qui rapportait plus de sensations de présence de manière générale que le reste des sujets. En exploitant uniquement leur résultats, André relève une moins grande sensation de présence pour les versions avec une importante cohérence audiovisuelle spatiale.

"*Remi Carreau et Thibault Macquart (2015) : Utilisation de la technologie WFS dans la création sonore cinématographique : possibilités et limites*", mémoire de fin d'études de l'École Nationale Supérieure Louis Lumière

Dans ce mémoire de fin d'études, Carreau et Macquart abordent les possibilités offertes par la technologie Wave Field Synthesis dans la création sonore au cinéma à travers un test perceptif comparant des séquences mixées sur un système 5.1 et sur un système WFS.

Les critères étudiés lors de cette expérience sont les suivants :
- La cohérence du mixage à l'image
- La profondeur et le relief
- La préférence

Ici les auteurs définissent la notion de cohérence de la manière suivante : "*Par ce critère de cohérence, nous entendons ici la fusion entre le champ visuel et le champ sonore, entre l'action vue et entendue, entre les sensations visuelles et sonores.*" La notion de cohérence employée ici est légèrement différente que celle développée dans ce mémoire. Dans cette étude on parlera de cohérence audiovisuelle spatiale pour décrire une cohérence physique entre le positionnement d'un son et de son image à l'écran. La définition de cohérence de Carreau et Macquart comprenant notamment des sensations de réalisme du rapport son/image correspond plutôt à un critère d'adéquation de l'espace sonore et de l'espace visuel.

Dispositif et Protocole : Ce test a été effectué par 20 sujets, tous sensibilisés au domaine du son au cinéma et répartis dans deux zones de la salle (centrée ou excentrée par rapport à l'écran).

Le test est composé de 6 extraits de courts métrages d'une durée allant de 1'05 à 3'50 pour la plus longue. Toutes les séquences sont diffusées via un système WFS. Le test se présente sous la forme d'une comparaison A/B. Les sujets voient dans un ordre aléatoire les six séquences. Pour chaque séquence ils se voient présenter un mixage 5.1 puis un mixage WFS (ou inversement). Une répétition dans un ordre différents est effectué.

Après chaque projection, les sujets doivent décider quelle version de la séquence est la plus cohérente à l'image, quelle version a le plus de profondeur et de relief, et quelle version ils ont préféré. Chaque échelle va de -5 à +5, les notes positives sont à l'avantage du mixage WFS. Le test dure en tout un peu plus de 55 minutes.

Résultats : Les expérimentateurs ont obtenu des résultats significatifs quant au critère de profondeur. En effet, pour 4 séquences sur 6 les sujets ont jugés que les mixages WFS apportait plus de profondeur et de relief.

En revanche aucun résultat probant ne se dégage quant aux critères de préférence, ou encore de placement dans la salle. De même, le critère de cohérence montre également des résultats non significatifs, et c'est ce qui nous intéressera plus particulièrement. Les sujets ne semblent pas avoir relevé de réelle différence de cohérence entre les différents mixages.

Discussion : On peut tenter d'expliquer l'absence de résultats significatifs concernant l'impact de la WFS sur la cohérence du son à l'image par plusieurs raisons. Tout d'abord il est possible que la définition de la cohérence du son à l'image par les expérimentateurs n'est pas été assez clair pour les sujets. De plus, comme les stimuli étaient très fournis en terme de matière sonore (hormis une séquence en intérieure où juste une voix est spatialisée en WFS) le jugement de cohérence a pu être compliqué à établir.

Ensuite, les séquences étaient très longues et le test par conséquent l'était aussi. Si les expérimentateurs écartent l'impact de la fatigue sur la durée totale du test, ils ne soulèvent pas la possibilité d'une sensation de lassitude au sein de chaque visionnage de séquence. Les sujets notent au minimum 2 minutes après le début du visionnage et parfois 8 minutes après. On peut imaginer que le sujets n'arrivent par à réellement statuer entre les deux versions. C'est peut-être notamment pour cela qu'aucune préférence ne se dégagent.

Cependant, des discussions avec les sujets ont montré que les scènes où les voix étaient spatialisées (dans la version WFS) ont fait réagir. Plusieurs sujets ont admis avoir été perturbés par la localisation précise de la voix. Si cela ne se fait pas nécessairement ressentir dans le résultats, il apparait intéressant de creuser ce sujet. Ces remarques semblent corroborer une hypothèse avancée par ce mémoire, à savoir que le déplacement de la voix de l'enceinte centrale peut être perçue comme dérangeant par les spectateurs.

2.2.2 Expérience V de la thèse de Etienne Hendrickx

Dans sa thèse "Cohérence des systèmes de diffusion sonore appliquée au cinéma en 2D et en 3D" (Université de Brest 2015) Etienne Hendrickx a effectué 5 expériences autour de l'impact de la stéréoscopie sur l'attente des spectateurs au niveau du son d'une séquence. Les trois premières expériences portent sur notre perception des sons d'ambiances pour un film en 3D. La quatrième porte sur l'effet ventriloque en élévation. Mais c'est l'expérience V qui nous intéressera plus particulièrement : Influence de la stéréoscopie sur l'appréciation de la cohérence audiovisuelle spatiale [7].

Description du test et des hypothèses

Cette expérience se base sur le fait que des chercheurs et des ingénieurs du son "*ont suggéré que la cohérence audiovisuelle spatiale pouvait améliorer significativement l'expérience des spectateurs, surtout pour des contenus en 3D*" (Hendrickx 2015). Aussi plusieurs ingénieurs du son pensent spatialiser différemment les sources pour les versions en 3D, notamment en les latéralisant.

De plus les quelques études ayant été menées n'ont pas réussi à s'accorder sur des résultats, quant à l'influence de la cohérence audiovisuelle spatiale, ou de l'impact de la 3D sur ce paramètre. C'est le cas des études d'André et al (2012) et de Moulins (2015), présentées plus haut, qui obtiennent des résultats contradictoires.

Il règne donc pour l'instant une ambiguïté quant à l'apport de la cohérence audiovisuelle spatiale, que Hendrickx explique notamment par le manque de variété des séquences utilisés lors des scènes. Le but de son expérience est de "*réévaluer la pertinence de la cohérence audiovisuelle spatiale au cinéma*" (Hendrickx, 2015). Dans son étude il étudie la cohérence audiovisuelle spatiale selon deux axes : en azimut et en profondeur.

Système de diffusion

La bande son des séquences était diffusée sur un système de diffusion composé de 7 enceintes Amadeus PMX4. Cinq enceintes étaient disposées derrière l'écran et deux autres étaient derrière le sujet et servaient de canal surround [8] .

Les stimuli

Le test présentait un total de huit séquences tournées nativement sur une caméra 3D. Le but est de couvrir le maximum de situations possibles, aussi la nature des sources sonores, la

7. Cette expérience a fait l'objet d'une publication : Hendrickx, E., Paquier, M. et Koehl, V. (2015). "Audiovisual spatial coherence for 2D and stereoscopic-3D movies" JAES 63 p.889-899, 2015

8. On retrouve cela dans différents systèmes de diffusions ayant disparus (SDDS, 70mm TODD-AO, Cinerama), ou qui voient le jour aujourd'hui (Dolby Atmos et prochainement NHK 22.2).

valeur des cadres, les décors, la dynamique des plans, varient selon les séquences. La bande son de chacune des séquences peut se séparer entre : des objets sonores (dialogue, voiture, effet sonore, etc...) et des ambiances couvrant l'espace 5.0 (LCRLsRs).

Différents mixages sont proposés pour chaque séquences. Le mixage des ambiances reste inchangé entre les différentes versions, et c'est uniquement la spatialisation des objets sonores qui change.

En azimut :
- Mixage "classique" : les objets sonores sont diffusés dans l'enceinte centrale
- Mixage "cohérent" : les objets sonores sont reproduits à l'azimut correspondant à la position à l'écran de leur image

En profondeur :
- Un mixage "proximité" : Les objets sonores sont mixés sans simulation de profondeur
- Un mixage "distance simulée" : Les objets sonores sont mixés en simulant une profondeur correspondant à l'image. Via des outils de mixage classiques (réverbération, équalisation, niveau) ou via l'utilisation de prises de son synchrones en champs diffus.

De même, chaque séquence étaient présentées dans sa version 3D et sa version 2D. Ainsi en croisant les différentes variables, chaque séquences étaient présentées selon 8 versions différentes. Les sujets devaient donc visionner 64 stimuli différents. Il est également à noter que pour chaque séquence, plusieurs objets sonores pouvaient être en présence, et donc chaque objet sonore étaient spatialisés de manière cohérente.

Protocole

Le test a été effectué par 16 sujets naïfs [9]. Après chaque stimulus, les sujets devaient répondre à la question suivante : "À quel point jugez-vous le son de cette séquence adapté à l'image ?". Afin d'y répondre il devait placer un curseur sur une échelle sans graduation allant de "très adapté" (100/100) à "pas adapté du tout" (0/100) [10]. Cette question permet d'obtenir un jugement global, plus écologiquement valable qu'une question focalisant l'attention du sujet sur la dimension spatiale, qui serait donc trop discriminante.

Les stimuli ont été présentés dans un ordre aléatoire pour chaque sujet. Une fois les 64 stimuli visionnés, le sujet faisait une pause de 15 minutes avant de refaire le test une deuxième fois, afin de tester l'impact de la répétition.

9. On parle de sujets "naïfs" dans ce cadre-là pour des gens qui ne sont pas en lien avec le cinéma ou le travail du son.
10. La même échelle est utilisée par Kruszielski et al (2012) et Kamekawa et al (2011).

Analyse des résultats

L'effet de la simulation de la profondeur a été jugé significatif par l'ANOVA et la moyenne pour les séquences avec simulation de profondeur est de 67 contre 64 sans traitement de la profondeur. Cependant en analysant les résultats, on se rend compte qu'il n'y a qu'une séquence où la simulation de la profondeur a réellement eu un effet sur le jugement des sujets [11]. Il est également à noter que le mode visuel (2D ou 3D) n'a eu aucune influence sur le jugement des sujets.

Le résultat le plus pertinent dans ce test reste l'influence de la cohérence en azimut. En effet, la moyenne des notes des séquences avec cohérence azimutale est de 71 contre seulement 60 pour des mixages "classiques". Il y a donc un réel effet de la cohérence azimutale dans le jugement d'adéquation du son à la séquence, qui est souligné par ce test.

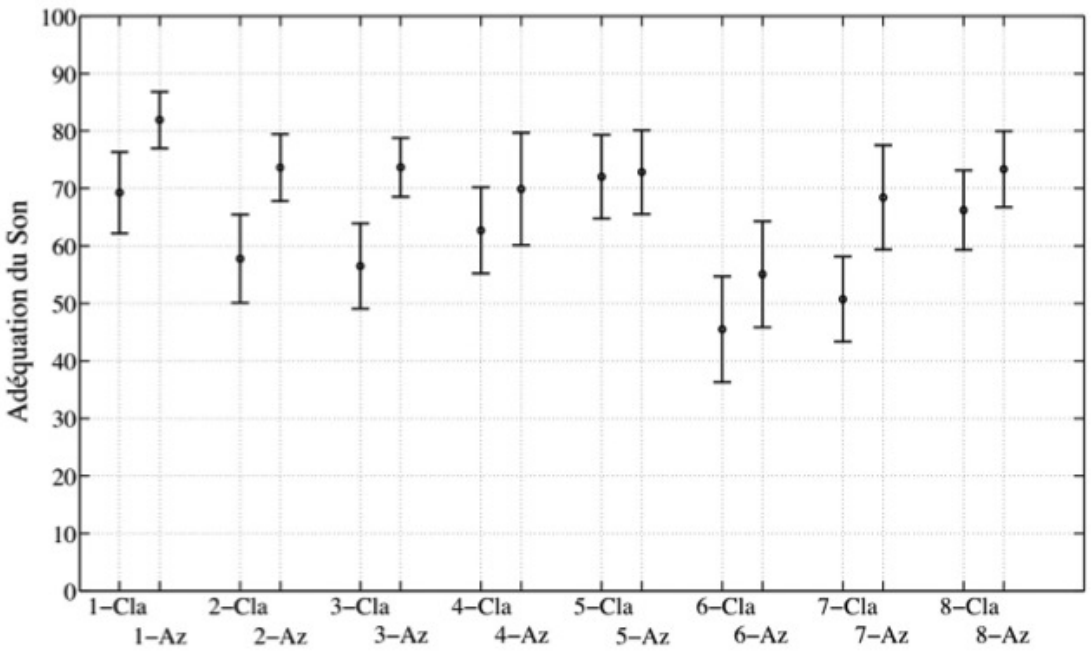


FIGURE 2.2 – *Effet de la cohérence Azimutale sur l'adéquation du son à l'image pour les 8 séquences. "Cla" = Mixage classique / "Az" = mixage cohérent en azimut. (Hendrickx, 2015)*

La figure 2.2 montre les résultats obtenus pour l'effet de la cohérence azimutale en fonction des huit séquences. On peut remarquer que l'impact de la cohérence azimutale est extrêmement dépendant de la séquence. En effet, si les séquences 1, 2, 3 et 7 montrent une amélioration importante, d'autres comme la séquence 8 sont plus ambiguës sur l'effet de la cohérence en azimut, voire il n'y en a aucun (séquence 5).

11. La séquence présentait un homme poussant un chariot à l'image et qui s'éloignait dans la profondeur.

Il est à noter également que la figure 2.2 présente une moyenne sur les deux sessions. Il s'avère que si l'on ne regarde que les résultats de la deuxième session, l'effet de la cohérence azimutale est plus marqué, et les écarts sont plus importants notamment pour les séquences 4 et 8 qui avaient des résultats ambigus sur la moyenne des deux sessions.

Hendrickx a donc cherché des corrélations entre les résultats et les paramètres d'azimut moyen et d'azimut maximal, ainsi que pour la vitesse moyenne et la vitesse maximale. C'est ce que présente la figure 2.3. Ici, les expérimentateurs ont fait le choix de ne traiter la corrélation en ne prenant compte que de l'objet sonore qu'ils considéraient comme le plus pertinent.

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

FIGURE 2.3 – *Récapitulatif des caractéristiques spatiales en azimut des objets sonores retenus pour chaque séquence du test, et l'amélioration apportée. (Hendrickx, 2015)*

Concernant l'azimut des objets sonores, il semble que plus les objets sonore ont un azimut moyen élevé, plus l'effet de la cohérence azimutale est important.

Les mêmes résultats sont observables quant à la vitesse moyenne qui semble avoir un impact sur l'influence de la cohérence azimutale. Cependant, plusieurs séquences présentant une vitesse moyenne nulle (seq 7 et 8) ont obtenu des résultats probant quant à l'amélioration de l'adéquation du son à l'image. Il semble donc intéressant d'aller plus loin dans l'étude des paramètres pouvant influencer le jugement des sujets.

2.3 Conclusion

Nous avons vu dans cette partie, que l'effet ventriloque permet au cerveau dans le cadre d'une projection audiovisuelle de corriger les différences de positionnement entre le son et l'image. Cependant, certaines études semblent suggérer malgré tout un réel apport de la cohérence audiovisuelle spatiale en azimut sur l'adéquation d'une bande son à une image.

Pour concevoir mon expérience à venir, il est important de prendre en compte les précédentes études :

- L'expérience de Hendrickx a indiqué que la cohérence audiovisuelle en profondeur ainsi que le mode de diffusion visuel (2D ou 3D) n'avait pas d'impact significatif sur le jugement des sujets.

- À l'inverse, la cohérence audiovisuelle spatiale en azimut semble être perçue comme une amélioration.

- De plus, l'amélioration apportée, semble être corrélée à l'azimut moyen des objets sonores et à leur vitesse moyenne.

- L'expérience de Carreau et Macquart, nous met en garde sur l'impact de la durée des séquences, et de bandes son trop chargées, pour réellement faire des différences pertinentes dans le jugement des sujets.

On peut donc définir les différentes hypothèses à vérifier pour notre expérience ainsi que les conditions auxquelles il faudra faire attention.

- Pour vérifier que la vitesse moyenne est bien corrélée avec l'amélioration apporté par la cohérence azimutale, il sera intéressant de faire varier le paramètre "vitesse" pour une seule et même source, afin de vérifier que c'est bien la vitesse et non la nature de la source sonore qui est importante.

- De plus, il semble judicieux de s'intéresser à l'impact que peut avoir la présence d'un montage image au sein d'une séquence, sur le jugement de la cohérence azimutale.

- Alors que Hendrickx ne fait aucune différence en fonction de la nature des objets sonores, il serait intéressant de vérifier si les résultats sont les mêmes quand l'objet sonore est une voix ou un effet. Ainsi, on pourra voir si la voix est jugée de la même manière que n'importe quel objet sonore dans une séquence cinématographique.

- Pour se rapprocher des conditions de l'expérience V de Hendrickx, on pourra reprendre l'échelle qui y est utilisée. La question de l'adaptation ou non du son à l'image peut paraître un peu trop orientée pour le critère étudiée (la cohérence audiovisuelle spatiale). Cependant, ce système semble avoir permis de mettre en évidence des résultats dans cette étude, et il serait donc plus rigoureux d'utiliser la même échelle de notation.

- De même, la répétition semble avoir eu un réel impact sur les résultats, il paraît donc intéressant de voir si les sujets s'habitueront à l'effet de déplacement de la voix.

- Pour éviter le problème de séquences trop longues pouvant fatiguer les sujets, il faudra utiliser des séquences assez courtes, autour de 20 secondes.

Ainsi, nous essaierons de mieux comprendre les mécanismes qui régissent la sensation d'adéquation à l'image par les spectateurs. De plus, nous tenterons de questionner la pertinence de la cohérence audiovisuelle spatiale pour la voix.

Deuxième partie

Contribution du mémoire

Chapitre 3

Réalisation d'un test perceptif sur la cohérence audiovisuelle spatiale en azimut de la voix

Le but de ce mémoire est d'explorer les possibilités et les limites de la cohérence de positionnement en azimut de la voix vis à vis de l'image. Claude Baiblé suggère que les mouvements frontaux des sons permettent de cultiver l'attention du spectateur, et donc sont perçus positivement [1] . L'expérience V de la thèse de Hendrickx semble confirmer que la cohérence de positionnement des objets sonores en azimut améliorerait l'expérience du spectateur dans de nombreux cas. Cependant, l'expérience portant également sur d'autres critères, à savoir l'impact de la stéréoscopie et d'un mixage cohérent en profondeur, elle ne creuse pas suffisamment certains aspects de la cohérence azimutale. De plus, la voix est traitée indifféremment des autres objets sonores, et chaque séquence présente des situations foncièrement différentes, tant sur la nature de la source, que sur la nature des placements (avec mouvement ou non/ proche ou non/ grande ou petite excursion en azimut).

Le but de cette expérience est donc de creuser la question de la cohérence en azimut des objets sonores et de voir si les spectateurs accorderont un traitement similaire à la voix et aux effets sonores.

Présentation succincte du test perceptif

Avant d'aller plus loin, rappelons rapidement les différents enjeux du test et les hypothèses qui ont mené à l'établissement de celui-ci.

Ce test a pour but d'étudier l'impact sur un spectateur que peut avoir la cohérence de placement d'un son par rapport à son image à l'écran, dans le cadre d'une diffusion cinémato-

1. Claude Baiblé, "L'image frontale, le son spatial" dans *Cinéma et dernières technologies* dirigé par Frank Beau, Philippe Dubois et Gerard Leblanc, INA et De Boeck et Larcier, 1998, p.243

graphique. Nous étudierons l'appréciation de la cohérence azimutale selon plusieurs critères, notamment l'impact du mouvement ou encore du montage image. De plus, nous chercherons à voir si le spectateur réagit de la même manière selon la nature de l'objet sonore. Pour la suite on séparera les objets sonores en deux catégories : les voix et les effets [2]. Différentes hypothèses motivent cette étude :

-> La cohérence audiovisuel en azimut peut améliorer la qualité d'expérience par le spectateur d'une séquence donnée.

-> La latéralisation de la voix ne sera probablement pas appréciée de la même manière que d'autre sources sonores présentes à l'image par le spectateur.

-> La latéralisation des sources de manière cohérente sera perçue plus positivement lorsque la source sonore sera en mouvement.

-> A contrario elle sera perçue plus négativement lorsque la séquence comprendra un montage image.

Afin de pouvoir tester ces hypothèses, il a donc été décidé de réaliser un test perceptif comportant 6 séquences. Trois séquences ont pour source sonore principale la voix, les trois autres sont centrées sur un effet sonore. Chaque séquence est proposée avec quatre situations différentes :

- 1 : La source sonore est fixe
- 2 : La source sonore est en mouvement lent
- 3 : La source sonore est en mouvement rapide
- 4 : La source sonore se déplace "instantanément" par le biais d'un découpage de la séquence.

Contrairement à la situation 4, les trois premières situations sont présentées sous la forme de plans séquences. Enfin, chacune sera proposée avec deux mixages différents. Le premier sera un mixage "classique", ou non-cohérent où la source sonore sera mixée dans l'enceinte centrale peu importe sa place à l'écran. Le deuxième sera un mixage cohérent, où la position en azimut de la source sonore correspondra à la position de son image à l'écran.

Ainsi nous avons un panel de situations plus important que dans l'expérience V de Hendrickx. Cela nous permettra d'envisager quels critères jouent un rôle dans l'influence de la cohérence en azimut sur la perception d'une séquence.

Les sujets verront dans un ordre aléatoire les séquences, et chaque séquence durera entre 10 et 40 secondes, afin que la durée du test reste raisonnable. Comme dans l'expérience de Hendrickx, les sujets devront juger s'ils jugent le son adapté ou non à la séquence.

2. Dans ce terme d'effet sonore on regroupe à la fois les effets réalistes proches du bruitages, et des effets moins réalistes.

3.1 Choix des séquences et tournage

3.1.1 Écriture des séquences

Afin de pouvoir tester les différentes situations pour une même source sonore il a été décidé très rapidement que les séquences seraient entièrement réalisées par l'auteur. Réaliser ces séquences m'a également permis d'approcher les problématiques posées par la cohérence azimutale sur la manière d'aborder le son direct.

Nous avons essayé de regrouper un ensemble de sources et de situations les plus diverses possibles. Ainsi, pour les trois séquences avec effets sonores, nous avons des sources continues (Mobylette et roller) plus ou moins large spectralement, ainsi qu'une source percussive et discontinue (le marteau). Les trois séquences avec de la voix présentent chacune une situation différente. Dans l'une on a des plans très large avec beaucoup de mouvements (le jogging), la séquence 5 nous permet de voir une situation qui suggèrent moins la notion de mouvement que l'action des joggeuses. Enfin, la dernière mettra en scène deux personnages spatialement séparés, avec un père et sa fille à deux endroits différents d'une pièce, nous testerons de même la situation du champ contre-champ.

En résumé

Ce test comprendra donc six séquences :

- Séquence 1 : Un homme fait du roller sur un parking
- Séquence 2 : Un homme fait de la mobylette
- Séquence 3 : Un homme cloue une planche
- Séquence 4 : Deux femmes parlent durant leur jogging
- Séquence 5 : Un homme est au téléphone avec sa mère
- Séquence 6 : Un homme dispute sa fille dans leur salon

Chaque séquence, et notamment leur contenu sonore, sera expliqué et précisé dans la suite de ce document. Par la suite on pourra se référer aux différentes séquences par leur numéro défini ci-dessus.

3.1.2 Matériel

Le matériel utilisé lors du tournage des séquences a été composé à la fois d'emprunt du matériel de l'école, de prêt et de location.

Matériel son

Pour ce tournage nous avons utilisé un enregistreur Sounddevices 664 et les microphones suivant : un Senheiser MKH50 (cardioïde), un Senheiser MKH416 (Semi-canon) et un couple MS Schoeps (deux corps de micro CMC6 associés à une capsule MK4 et une capsule MK8). S'ajoute à cela deux ensembles HF audiolimited 2020 accompagnés de microphones cravates Sanken Cos 11.

Il était indispensable pour le tournage de disposer de systèmes HF, et d'un enregistreur présentant suffisamment d'entrées. En effet, il était nécessaire d'avoir une prise de son la plus propre possible sur les voix, afin de pouvoir les séparer si nécessaire de l'environnement sonore alentour. De plus, pour les besoins de mes tests de nombreux cadres étaient très large, ce qui empêchait de capter les voix de manière acceptable (notamment pour la scène de jogging).

J'ai fait le choix de tourner tous les plans avec un couple MS sur pied. L'idée était avant tout de pouvoir lors de la post-production me confronter au problème qu'un ingénieur du son pourrait rencontrer lorsque des ambiances synchrones ou des couples synchrones ont été utilisés lors du tournage. Ainsi, le couple qui capte à la fois l'objet sonore qui m'intéressera pour le test, capte également tout l'environnement sonore, et surtout il place déjà l'objet sonore dans un espace en deux dimensions. Le but sera de voir si l'utilisation de ce couple synchrone sera compatible avec le placement cohérent des objets sonores à l'image.

Le couple MS n'est peut-être pas le couple qui a le meilleur rendu sur les timbres et l'espace sonore (notamment dû à l'utilisation d'un microphone bidirectionnel), mais il offre une grande marge de manœuvre en post-production. De plus, c'est un couple qui est assez souvent utilisé dans le cadre de tournage de fiction.

3.1.3 Méthodologie de tournage

L'enjeu principal des tournages était de faire en sorte d'avoir la matière la plus propre et la plus complète possible. Aussi pour pouvoir plus facilement spatialiser l'objet sonore dont il est question dans la séquence, il faut pouvoir séparer dès la prise de son le plus possible les différents éléments.

Aussi dès que possible des micros cravates ont été utilisés, y compris sur les rollers et la mobylette. Il n'y a que la séquence 3 qui n'utilise pas de système HF, cela n'étant pas utile au vu du niveau de la source sonore (un marteau).

De même le MKH50 a très souvent servi de micro d'appoint caché dans le décors. on retrouve cette méthode dans la séquence 1 (roller) , la séquence 5 (téléphone) et la séquence 6 (dispute père-fille).

Sur chaque décor, des ambiances ont été enregistrées ainsi que des silences raccords dans les lieux plus calmes. Pour une description plus détaillée des tournages vous pouvez vous reporter aux annexes.

3.2 Post-production des séquences et premières hypothèses

La partie qui suit présente succinctement la post-production des séquences, vous trouverez des explications plus détaillées concernant le montage des ambiances et le mixage en annexe.

Avant d'aborder plus en détails le montage son de l'ensemble de ces séquences, il est à noter que j'ai effectué également la post-production image. J'ai donc dérushé et sélectionné les plans présentant le meilleur compromis entre mon sujet et une séquence efficace. En effet, je me suis confronté à certains problèmes liés aux rushes images. Certains plans qui me semblaient mieux au tournage du fait du jeu et d'une esthétique plus intéressante, n'étaient finalement pas forcément les plus adaptés pour mon expérience. Certains présentaient trop peu de mouvement car ayant été instinctivement trop suivis par la caméra, où bien les différences de vitesses entre la version lente et rapide du mouvement n'étaient pas assez marquées sur les plans initialement choisis. Il a donc fallu faire des concessions, au niveau de la qualité de jeu ou de la perfection à l'image de certains plans, le choix s'est porté naturellement vers les plans se prêtant le plus possible à mon expérience, au détriment du jeu.

3.2.1 Montage son des séquences

Le montage direct

Comme lors du tournage des séquences, le défi était ici de préparer au mieux le terrain pour permettre un placement des sources sonores qui ne choquerait pas et qui soit le plus propre possible. Il faut donc impérativement que le montage des directs soit impeccable.

Pour la grande majorité des séquences le travail consistaient avant tout à choisir les prises les plus adaptées et présentant le meilleur compromis qualité de son et jeu, faire le tri dans les différentes pistes et muter les pistes inutiles, ainsi qu'effectuer la remise en phase perche-HF quand cela était nécessaire. J'ai également proposé pour différentes scènes des versions dénoisées des directs pour préparer le terrain pour le mixage. Les cinq premières séquences ne présentant pas de cas problématiques vis à vis du sujet de l'expérience ne dérogèrent pas aux cas classiques.

La séquence 6 cependant fut plus intéressante concernant le montage direct. En effet, cette scène présente deux individus en train de discuter. Cependant contrairement à la séquence 4 où les joggeuses sont situées au même endroit à l'image, les deux personnages sont en permanence à des placements différents. Cela implique donc que pour le mixage cohérent nous aurons deux objets sonores : le père d'un côté et la fille de l'autre. Lorsque le père parle, le micro placé sur la fille capte un champ diffus du discours du père et inversement. Si cela ne serait pas nécessairement gênant dans un mixage "classique", ce peut devenir problématique pour une séparation spatiale entre la fille et son père. Lors du tournage, des silences plateaux ont été enregistrés en respectant le placement des acteurs, pour avoir des silences raccords correspondant vraiment aux fonds d'air captés par les deux micros utilisés sur chaque comédiens. Ces silences ont donc été utilisés pour "boucher" les moments d'inactivité de chaque personnages. Cela permet donc d'obtenir une continuité sonore pour chaque objet sonore sur l'ensemble de la séquence et de les rendre ainsi indépendants l'un de l'autre, permettant donc une spatialisation également indépendante.

Le montage des ambiances

Pour se rapprocher le plus possible de séquences de cinéma classique, un montage d'ambiance et d'effet a été réalisé en 5.0. Afin de ne faire varier que le mouvement de la source sonore, le montage des ambiances reste globalement le même pour les différentes versions d'une même séquence. Les sons constituant ces ambiances provenaient à parts égales de sonothèques et d'enregistrements effectués directement par l'auteur, sur les lieux de tournages des séquences, ainsi que sur des tournages précédents. Ces enregistrements ont été effectués pour la grande majorité en MS ou en DoubleMS.

En plus des ambiances alimentant les canaux gauche, droite et surrounds, des ambiances monophoniques sont diffusées en permanence dans le canal central. Cela permet de garder un équilibre lors des déplacements des objets sonores, mais aussi de lier cet objet au reste de l'espace sonore.

3.2.2 Expérience de mixage

L'ensemble des séquences a été mixé dans l'auditorium de mixage de l'ENS Louis Lumière. Afin d'optimiser le déplacement des objets sonores, il a été décidé de travailler avec cinq enceintes derrière l'écran, comme sur l'expérience de Hendrickx. Il n'était donc pas possible d'utiliser le système Meyer déjà installé. Cinq enceintes APG DX8 ont donc été utilisées pour nourrir les cinq canaux frontaux.

Comme il y avait 24 séquences à mixées et que de plus la spatialisation des sources sonores se faisait en partie en différé, le choix a été fait de ne pas mixer en sortant les pistes

sur la console Euphonix System 5 mais d'utiliser le protocole EUCON permettant un contrôle de la session et des pistes depuis la console.

Plutôt que d'utiliser un bus au format 7.1 SDDS, comme le propose Protools HD, il a été décidé de spatialiser les objets sonores par le biais d'un patch max/Msp. Ce patch permet de spatialiser un son monophonique sur cinq enceintes (grâce à une loi tangentielle classique) et prend de plus en compte la position des enceintes. Cette méthode se fait obligatoirement en différé, et n'est donc pas très orthodoxe. Cependant, cela nous permet de nous affranchir de l'erreur humaine que peut apporter un mixage à "l'oreille".

Premières impressions

Durant le mixage j'ai pu me faire mon avis sur ce qui marche ou ce qui ne marche pas en terme de mixage cohérent. Les quelques remarques qui suivent reflètent uniquement une impression de l'auteur sur le mixage des séquences.

De manière générale, pour les mixages cohérents, il m'a semblé que le regard était orienté par le son. L'image fournit déjà des informations sur le placement des éléments constituant la séquence, et oriente nécessairement le regard en fonction de ce qui est mis en scène. Cependant, l'apport du son cohérent me semble un moyen de diriger encore plus le spectateur vers un point de l'image.

De plus, ces effets sont souvent associés à des images présentant des objets sonores éloignés du centre de l'image, et souvent en plan assez large. L'effet de décentrement est donc déjà très fort à l'image, il n'est donc pas forcément illogique de l'accompagner par un décentrement du son, et cela paraît même aider à orienter le spectateur.

Le cas de la séquence 6 (le père et sa fille) me semble intéressant, du fait que les deux objets sonores soient séparés dans l'espace (et symboliquement dans l'histoire). En effet, le fait de placer leur voix de manière cohérente à l'image, renforce cette séparation entre le père et sa fille. Mais c'est aussi cette séquence qui semble la plus enclin à être jugée négativement par les sujets du test. Le fait de passer instantanément de la voix du père sur la gauche, à celle de la fille sur la droite me semble un peu trop artificiel et donc susceptible d'être gênant, car sortant le spectateur de la narration.

Sur les mouvements : À la première écoute, le suivi des mouvements au son, et plus particulièrement sur la séquence des rollers ou encore celle de la mobylette, semble réellement apporter du dynamisme au mouvement déjà présent à l'image. La version rapide de la séquence 1 semble beaucoup plus appropriée avec le mixage cohérent, car à l'image le jeune homme va très vite et passe très proche de la caméra, donnant un effet de surgissement, qui marche assez peu avec le mixage classique.

Cependant, différents problèmes semblent être soulevés par le mouvement des sources sonores. Premièrement, il semble que le moindre défaut de placement du son par rapport à l'image ressortent beaucoup, en effet si le son suit constamment l'image, le moindre écart devient gênant. Alors même que lorsque tout est mixé au centre, cela ne pose pas réellement de problème.

De plus, les mouvements des sons soulignent le problème lié aux réglages des enceintes. En effet, pour que l'effet soit parfait, il faudrait que les enceintes correspondant aux cinq canaux frontaux utilisées, soient parfaitement identiques. Il est donc nécessaire de corriger tant les différences de niveaux que les différences de timbres. Cependant, les différences de timbres sont difficiles à maitriser totalement, et il est difficile d'obtenir un réglage d'enceinte parfait. De ce fait, quand un son passe d'une enceinte à une autre, on peut avoir un léger changement de timbre qui peut être dérangeant pour les initiés.

Enfin, les mouvements des voix peuvent sembler étrange à une première écoute. Si les mouvements de la voix sur la séquence 4 paraissent assez adaptés du fait des plans très larges et du mouvement très présent des personnages à l'image, c'est moins le cas sur les deux autres séquences parlées.

- Sur la séquence du jeune homme au téléphone, si cela n'est pas réellement gênant sur le mouvement lent, voir agréable à l'écoute, cela ne semble pas avoir un réel apport pour la séquence. De plus, pour le mouvement rapide, l'effet paraît presque trop appuyer le mouvement déjà présent à l'image.

- Pour la séquence 6, les mêmes remarques s'appliquent pour le mouvement rapide. De plus sur le mouvement lent, le personnage porte une veste qui fait énormément de bruit. Si ce bruit ne dérange pas quand le son est au centre, on y fait beaucoup plus attention quand la voix bouge, ce qui devient alors dérangeant.

Sur les montages : La cohérence azimutale semble finalement pour la plupart des séquences asez adéquat. Pour les séquences 4 et 5, le mixage cohérent semble être assez approprié avec le montage image. Cela peut s'expliquer :

- Pour la séquence 4, bien que l'on passe de l'extrême droite à l'extrême gauche à deux reprises, cela s'accompagne d'abord d'un mouvement très marqué des personnages (elles courent) mais aussi d'un effet de surgissement des personnages, qui je pense, fait que le suivi en azimut fonctionne assez bien.

- Pour la séquence 5, je pense que ce qui fait que l'effet fonctionne plutôt bien, est que le personnage ne parlent pas très vite et laisse un peu d'espace entre ses répliques (il est au téléphone), cela permet donc de placer les points de montage image entre deux phrases, ce qui rend le déplacement de sa voix moins gênant.

Deux séquences semblent cependant poser problème :

- La première est la séquence 1, en effet le montage du roller ne perd jamais en intensité et on a donc des sautes en azimut importantes alors que l'intensité sonore est également importante, cela crée un sentiment de gêne, bien que cela renforce le dynamisme de la scène.

- La seconde est la séquence 6. Pour cette séquence j'ai décidé d'effectuer un champ-contrechamp entre les deux personnages, afin de tester les limites de la cohérence audiovisuelle spatiale. Les déplacements très fréquents de la voix du père sont assez gênants, surtout lorsque sa voix se retrouve à l'extrême de l'image. De plus, les coupes à l'image ont été intentionnellement placées au milieu de phrases, afin de nuancer avec la séquence 5, et c'est je pense un des facteurs qui peut déranger le spectateur.

3.2.3 Hypothèses quant aux résultats de l'expérience

Après le mixage des séquences et avant de passer à la mise en pratique de mon test perceptif plusieurs hypothèses semblent se dégager, quant aux résultats probables de l'expérience :

- De manière générale l'apport de la cohérence sur le mixage sera ressentie de manière positive pour les séquences ne mettant pas en scène de la parole.

- Les séquences avec un montage pour ces trois premières séquences poseront éventuellement problème, surtout le montage de la séquence 1 qui risque d'être noté moins bien pour le mixage cohérent (pour les raisons avancées plus haut).

- Pour les séquences vocales, on notera probablement un apport positif de la cohérence azimutale pour les versions sans mouvements, ou avec mouvement lent.

- La séquence 6 posera sans doute problème car présentant les limites du dispositif avec deux personnages placés différemment, notamment sur la version montée.

- La séquence des joggeuses sera probablement mieux notée sur l'ensemble de ses versions pour le mixage cohérent car étant la séquence se prêtant le mieux à l'effet.

Si certains résultats me semblent néanmoins dur à imaginer, je pense que concernant les séquences avec de la voix, il y aura soit une réaction foncièrement positive, soit une réaction totalement négative. Mais je ne m'attends pas à l'absence d'impact du déplacement de la voix.

3.3 Description du test perceptif

Avant d'aborder les résultats du test perceptif, nous allons effectuer un descriptif de l'expérience et des stimuli utilisés.

3.3.1 L'installation technique

Le test s'est déroulé dans l'auditorium de mixage de l'école Louis Lumière. Il s'agit de la même salle où les séquences ont été mixées. Le système de diffusion utilisé était le suivant :

- En façade, cinq enceintes APG DX8 pour les canaux L(Gauche), Lc(Inter-Gauche), C(Centre), Rc(Inter-Droite) et R(Droite). Les enceintes sont amplifiées par deux amplificateurs SA20 de APG.

- Les canaux arrières sont diffusés sur les enceintes JBL installées dans l'auditorium.

Le système a été calibré avant de mixer les séquences.

FIGURE 3.1 – *Installation des cinq enceintes APG DX8 diffusant les cinq canaux frontaux. (Les deux enceintes en bas à droite ne sont pas utilisées dans l'installation présente)*

Comme on peut le voir sur la figure 3.1, les enceintes sont posées sur une structure en métal montée pour l'occasion. Cela n'est pas nécessairement la meilleure solution pour installer des enceintes, mais elle permet de répondre aux problématiques de mon sujet mais aussi de l'expérience menée par Simon Prieur dans son mémoire sur un système de diffusion utilisant la verticalité. Nous avons donc mutualisé l'installation le temps de nos mixages et expériences respectives.

Le sujet du test était installé sur un siège de cinéma situé devant la console de mixage, de sorte qu'il n'aie devant lui que l'écran de projection. L'image était projetée sur une largeur de 4 mètres. Afin de retrouver des conditions similaires à l'expérience effectuée par Hendrickx dans sa thèse, il faut que l'ouverture du champ de vision soit à peu près de 46˚. Les sujets ont donc été placés à 4m80 de l'écran ce qui donne un angle proche de 45˚ [3].

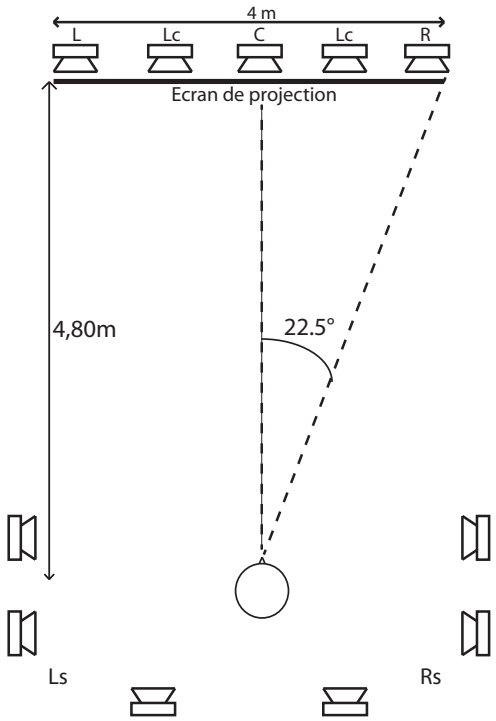


FIGURE 3.2 – *Position du sujet par rapport à l'écran de projection et des enceintes de diffusions frontales.*

Le niveau de diffusion a été décidé arbitrairement par l'expérimentateur avant les tests, comme c'est souvent le cas lors de test perceptif.

3.3.2 Le protocole

Treize sujets âgés de 19 à 61 ans ont pris part au test perceptif. Tous les sujets sont des naïfs, n'ayant pas connaissance du sujet de l'expérience et ne travaillant pas dans le milieu du son ou même du cinéma de manière générale. La fréquentation de salle de cinéma des sujets va de 3 à 6 fois par an pour les moins assidus à plus d'une fois par semaine pour certains.

Chaque sujet passe le test seul. Les sujets se voient présenter 24 séquences selon deux mixages différents :

- Un mixage cohérent, où les objets sonores suivent leur correspondant visuel en azimut.

- Un mixage classique, où les objets sonores sont diffusés uniquement par l'enceinte centrale.

3. C'est l'ouverture du champ de vision recommandée par Dolby.

Les 48 stimuli sont diffusés dans un ordre complètement aléatoire et différent pour chaque sujet pour éviter l'effet d'ordre. À la fin de chaque stimulus le sujet doit juger l'adéquation du son de la séquence à l'image. Les résultats sont enregistrés par le biais d'un patch Max/MSP (voir figure 3.3). Une fois le sujet satisfait de la note attribuée, il peut cliquer sur le bouton "suivant", pour voir la séquence suivante.

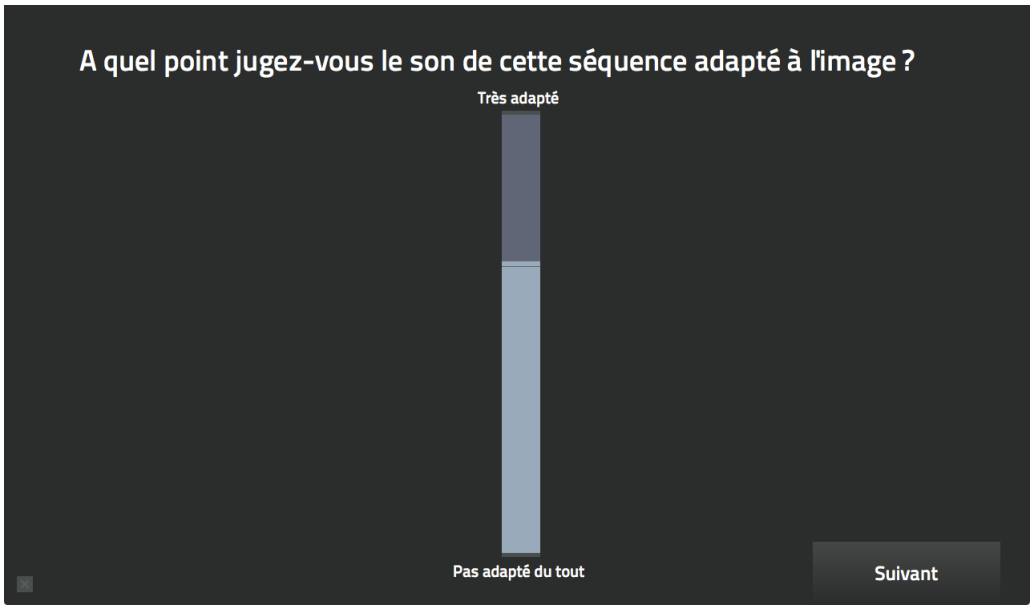


FIGURE 3.3 – *Interface de notation du test perceptif*

Avant le début du test, il est dit au sujet que ce que l'expérimentateur entend par "adapté" revient à juger le fait que le son "marche" bien avec l'image. Il leur est indiqué que le jugement doit se faire de manière global et spontané. Pour ne pas influencer les résultats, il ne leur est à aucun moment mentionné les paramètres évalués par le test, ni le terme de "cohérence".

Durée du test

Une fois le sujet installé, on réalise un pré-test d'à peu près 5 minutes. Durant ce pré-test, le sujet se voit présenter aléatoirement une version de chaque séquence. À la fin de chaque séquence il doit les noter. Cela permet au sujet de découvrir les six objets sonores différents, ainsi que de voir des séquences où le mixage est cohérent et d'autres où il est classique. De plus, cela lui permet de prendre en main l'outil de notation. Les notes pour ces six versions ne sont pas conservées, et une fois les six séquences visionnées, on revient à zéro et on commence une première session de test.

Une fois la première session terminée (c'est à dire une fois que les 48 stimuli ont été notés) le sujet est invité à prendre une pause en dehors de l'auditorium, où il peut boire et manger. Après 5 à 10 minutes de pause, on reprend le test pour une deuxième session, où le

sujet note à nouveau les 48 stimuli dans un ordre différent que pour la première session. Cela permet de voir comment évoluent les résultats avec le temps.

Chaque stimulus dure de 12 secondes pour le plus court, jusqu'à 40 secondes pour le plus long, avec une durée moyenne d'une vingtaine de secondes. La première session et le pré-test durent entre 35 et 40 minutes selon les sujets et la deuxième entre 25 et 30 minutes, pour une durée totale de 1h05 à 1h15 en comptant la pause. Le test peut sembler un peu long, c'est pourquoi les sujets sont invités à ne pas négliger la pause.

À la fin du test, il leur est demandé de remplir un questionnaire permettant de rapporter leurs impressions, et notamment de dire s'ils ont ressenti de la fatigue durant le test. Le temps de remplir le questionnaire est également un temps pour discuter avec le sujet et essayer de comprendre la manière avec laquelle ils ont appréhendé les différentes séquences, et ce qui les a gênés ou non.

3.3.3 Descriptif des stimuli

L'expérience V de la thèse de Hendrickx soulevait divers critères qui méritaient d'être approfondis. Aussi Hendrickx soulignait que l'amélioration apportée par la cohérence azimutale semblaient fortement dépendante à la fois de l'azimut moyen de l'objet sonore spatialisé mais aussi de sa vitesse moyenne. Cependant, l'expérience ne testait pas plusieurs vitesses pour une même source. Il semble donc intéressant de voir si les corrélations entre la vitesse et l'amélioration apportée se vérifient quand on a des variations de vitesses pour un même objet. Nous rajoutons également une condition avec un montage image qui correspondrait à une vitesse "infinie", par les sautes d'azimut provoquées via les coupes à l'image, cela permettant de voir les limites éventuelles de la cohérence azimutale.

De plus, l'expérience de Hendrickx ne faisait pas de différence sur la nature de l'objet sonore. Les résultats ne semblaient pas mettre en évidence de différences selon que l'objet sonore soit de la voix ou non. En outre, les scènes présentaient pour la grande majorité plusieurs objets sonores spatialisés, indifférenciés. Les corrélations entre l'amélioration apportée par la cohérence azimutale et la vitesse moyenne ou l'azimut moyen se faisaient en ne choisissant que l'objet sonore qui était selon l'expérimentateur le plus susceptible de créer une différence. C'est pourquoi ici, nous ne mélangerons pas la nature des sources sonores dans une même scène, et hormis la séquence 6 présentant deux objets complètement séparés dans l'espace, les autres séquences ne présenteront qu'un seul objet sonore spatialisé. Cela a pour but de vérifier la corrélation avec la vitesse retrouvée par Hendrickx, et de voir si cela se confirme peu importe la nature de l'objet sonore.

Les vitesses : Les 24 scènes différentes sont en fait, comme dit en introduction de cette partie, séparées en 6 séquences principales, elles-mêmes présentées avec quatre vitesses de

déplacement des sources différentes :

- Vitesse 1 : la vitesse est nulle, et on a une source sonore qui est fixe à l'image.
- Vitesse 2 : La source sonore est en mouvement lent.
- Vitesse 3 : La source sonore est en mouvement rapide.
- Vitesse 4 : La source a une vitesse "infinie", se déplaçant grâce à un montage image.

Cohérence azimutale : Comme dit précédemment, nous ne testerons ici que la cohérence audiovisuelle spatiale en azimut. En effet, plusieurs études ont montrés que l'effet ventriloque était extrêmement fort en élévation (Thurlow et Jack 1973, Hendrickx 2015). Concernant la cohérence en profondeur, Hendrickx n'a soulevé aucun impact sur le jugement de ses sujets. Chacune des 24 scénettes sera donc présentée avec deux mixages, un mixage classique et un mixage cohérent en azimut (Cf 3.3.2).

Séquence	Description
Séquence 1 : 	**Fixe :** Le jeune homme fait du roller au loin, il fait quasiment du surplace. **Mouvement lent :** Le jeune homme commence à rouler dans toute la largeur de la route assez lentement. **Mouvement rapide :** Le plan est plus serré, le jeune homme prend de la vitesse et passe tout près de la caméra à plusieurs reprise. **Montage image :** D'abord un plan large, puis un plan rapproché sur les rollers qui accélèrent, mouvement rapide et fin sur le plan fixe.
Séquence 2 : 	**Fixe :** Le jeune homme essaie de faire démarrer la mobylette sur le bord de la route. **Mouvement lent :** La mobylette arrive lentement dans la profondeur. **Mouvement rapide :** La mobylette fait entrée dans le champ par la droite, traverse l'image et sort du champ à gauche. **Montage image :** On commence sur le plan fixe, puis un plan rapproché sur le visage du jeune homme qui s'en va, puis le plan avec mouvement rapide.
Séquence 3 : 	**Fixe :** Plan moyen, le jeune homme enfonce un cloue dans une planche en bois. **Mouvement lent :** Même action, on a un plan serré, sur la planche, la caméra fait un mouvement latéral, le cloue entre et sort du champ. **Mouvement rapide :** Plan serré avec mouvement de caméra à l'épaule **Montage image :** Alternativement plan large, gros plan sur les clous, plan sur le visage, plan sur le marteau.

FIGURE 3.4 – *Descriptif des différentes versions des séquences avec objets sonores.*

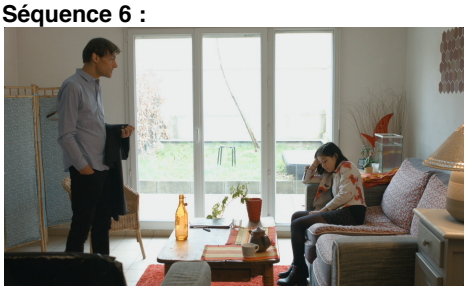
Séquence	Description
Séquence 4 : 	**Fixe :** Plan large, les joggeuses sont à l'arrêt et discute avant de partir faire leur footing. **Mouvement lent :** Les joggeuses entrent dans le champ par la gauche et marchent essoufflées jusqu'à sortir du champ. **Mouvement rapide :** Le plan est plus large, les joggeuses entrent en courant par la gauche la caméra les suis, elles sortent par la droite. **Montage image :** D'abord un plan sur les joggeuses qui viennent vers la caméra, puis le mouvement rapide, puis plan de dos sur les joggeuses qui s'éloignent.
Séquence 5 : 	**Fixe :** L'homme est adossé à la porte de sa véranda et téléphone à sa mère. **Mouvement lent :** L'homme sort de sa véranda et marche dans son jardin au téléphone. **Mouvement rapide :** Le plan est plus serré, l'homme sort dans son jardin et marche rapidement, il est énervé. **Montage image :** On commence sur le plan fixe, puis l'homme appelle sa femme, on a un plan de l'intérieur de la véranda, puis plan serré sur le visage de l'homme.
Séquence 6 : 	**Fixe :** Plan en pied du salon, de profil. Le père est debout à gauche et dispute sa fille assise sur le canapé à droite. **Mouvement lent :** Même valeur de plan, on est derrière le canapé. La fille est au centre, le père fait des vas et vient devant le canapé. **Mouvement rapide :** Plan en légère plongée derrière le père qui marche plus nerveusement. La fille est à gauche. **Montage image :** D'abord plan fixe assez large, puis champ/contre-champ.

FIGURE 3.5 – *Descriptif des différentes versions des séquences avec voix.*

Résumé des stimuli

Version de la séquence	Vitesse de la source sonore	Type de mixage
1	Fixe	Mixage "classique"
2	Mouvement lent	Mixage "classique"
3	Mouvement rapide	Mixage "classique"
4	Montage image	Mixage "classique"
5	Fixe	Mixage "cohérent"
6	Mouvement lent	Mixage "cohérent"
7	Mouvement rapide	Mixage "cohérent"
8	Montage image	Mixage "cohérent"

TABLE 3.1 – *Les huit déclinaisons possibles pour une séquence présentant un même objet sonore*

3.4 Exploitation des résultats

Le test a été effectué avec 13 sujets. Nous avons analysé les résultats grâce à une ANOVA (ANalyse Of VAriance) à mesure répétée. L'ANOVA est en général utilisée sur un plus grand nombre de sujets, mais plusieurs études ont considéré que l'ont pouvait accepter une ANOVA à partir de 12 sujets. Avant d'effectuer l'ANOVA, les résultats ont été normalisés afin de minimiser les différences d'utilisation de l'échelle par les sujets.

Une analyse sujet par sujet à amener à écarter le sujet n˚5. En effet, ce dernier adoptait régulièrement une stratégie différente de celle des autres sujets. Après discussion avec le sujet il semblerait qu'il n'ai pas exactement compris la question posée et il a donc systématiquement donné la note maximale aux mixages cohérents et une très mauvaise note aux mixages classiques, peu importe la vitesse ou la nature des objets sonores. Il a en fait cru devoir jugé à quel point le son était spatialement cohérent et n'a donc pas réellement répondu à mon attente d'avoir un jugement global.

Les résultats de l'ANOVA sont présentés dans le tableau 3.2. Les différents facteurs de l'analyse sont les suivants :

- R : Répétition (2 niveaux)
- Az : Cohérence Azimutale : mixage "classique" ou "cohérent" (2 niveaux)
- N : Nature de l'objet sonore : "effet" ou "voix" (2 niveaux)
- S : Séquence (3 niveaux) [4]
- V : Vitesse, fixe/ lent/ rapide/ montage image (4 niveaux)

Pour qu'une interaction entre les différents facteurs soit considérée comme significative, il faut avoir le facteur de significativité $Sig.p \leq 0.05$.

Ici, on peut remarquer que plusieurs résultats diffèrent de l'expérience de Hendrickx. En effet, le facteur Az n'est pas significatif ce qui voudrait dire que sur l'ensemble des stimuli, la version du mixage n'a pas eu d'effet sur les résultats. De même, l'interaction Az*R n'est pas significative, ce qui veut dire que la répétition n'a pas eu d'impact sur l'appréciation de la cohérence par les sujets, contrairement à l'expérience de Hendrickx.

Cependant, il semble que la nature de l'objet sonore aie eu une influence sur le jugement des sujets. Nous allons donc explorer la corrélation entre la nature des objets sonores et l'amélioration de l'adéquation du son à l'image.

4. Ici il s'agit de 3 séquences par nature d'objet différentes. Une autre analyse ANOVA a été faite sans prendre le facteur Nature et avec donc 6 niveaux pour le facteur Séquence.

Source	Somme des carrés	DDL	Moyenne des carrés	D	Sig.p
R	3625,082	1	3625,082	4,782	0,051
Az	3447,598	1	3447,598	1,062	0,325
N	3479,068	1	3479,068	3,426	0,091
S	3877,086	2	1938,543	4,171	0,029
V	6879,557	3	2293,186	6,212	0,002
R*Az	68,470	1	68,470	0,159	0,697
R*N	11,108	1	11,108	0,093	0,766
R*S	65,353	2	32,677	0,195	0,824
R*V	755,173	3	251,724	0,725	0,545
Az*N	3062,652	1	3062,652	8,249	0,015
Az*S	1224,419	2	612,210	1,329	0,285
Az*V	292,640	3	97,547	0,288	0,833
N*S	16,129	2	8,064	0,016	0,984
N*V	2308,209	3	769,403	1,241	0,311
S*V	4381,562	6	730,260	0,768	0,598
R*Az*N	13,326	1	13,326	0,050	0,827
R*Az*S	418,042	2	209,021	0,604	0,556
R*Az*V	1004,821	3	334,940	1,301	0,290
R*N*S	243,881	2	121,940	0,337	0,718
R*N*V	1078,714	3	359,571	1,063	,378
R*S*V	2330,468	6	388,411	1,621	0,155
Az*N*S	2128,633	2	1064,316	2,943	0,074
Az*N*V	2911,501	3	970,500	3,069	0,041
Az*S*V	2248,553	6	374,759	1,214	0,310
N*S*V	11165,195	6	1860,866	2,866	0,015
R*Az*N*S	900,748	2	450,374	2,392	0,115
R*Az*N*V	1028,789	3	342,930	1,330	0,281
R*Az*S*V	906,025	6	151,004	0,664	0,679
R*N*S*V	1563,125	6	260,521	1,118	0,361
Az*N*S*V	429,536	6	71,589	0,285	0,942
R*Az*N*S*V	730,133	6	121,689	0,466	0,831

TABLE 3.2 – *Résultats de l'ANOVA*

3.4.1 Influence de la Nature de l'objet sonore

Les interactions de facteurs Az*N et Az*N*V sont significatives, avec respectivement $p = 0.015 < 0.05$ et $p = 0.041 < 0.05$. Cela implique que l'appréciation du mixage cohérent par rapport au mixage classique n'est pas la même selon la nature de l'objet sonore. De plus, cela semble dépendre également de la vitesse de ces objets.

La figure 3.6 semble clairement montrer que les sujets n'ont pas jugé de la même manière les séquences en fonction de la nature des stimuli. Pour le mixage "classique", les résultats semblent globalement les mêmes peu importe que l'objet soit de la voix ou non.

Cependant pour le mixage "cohérent", alors que pour les objets la cohérence azimutale semble avoir eu un impact réellement positif sur le jugement d'adéquation par les sujets, cela ne semble avoir eu aucune incidence pour les séquences avec de la voix. Cela explique en partie l'absence de significativité pour le facteur Az.

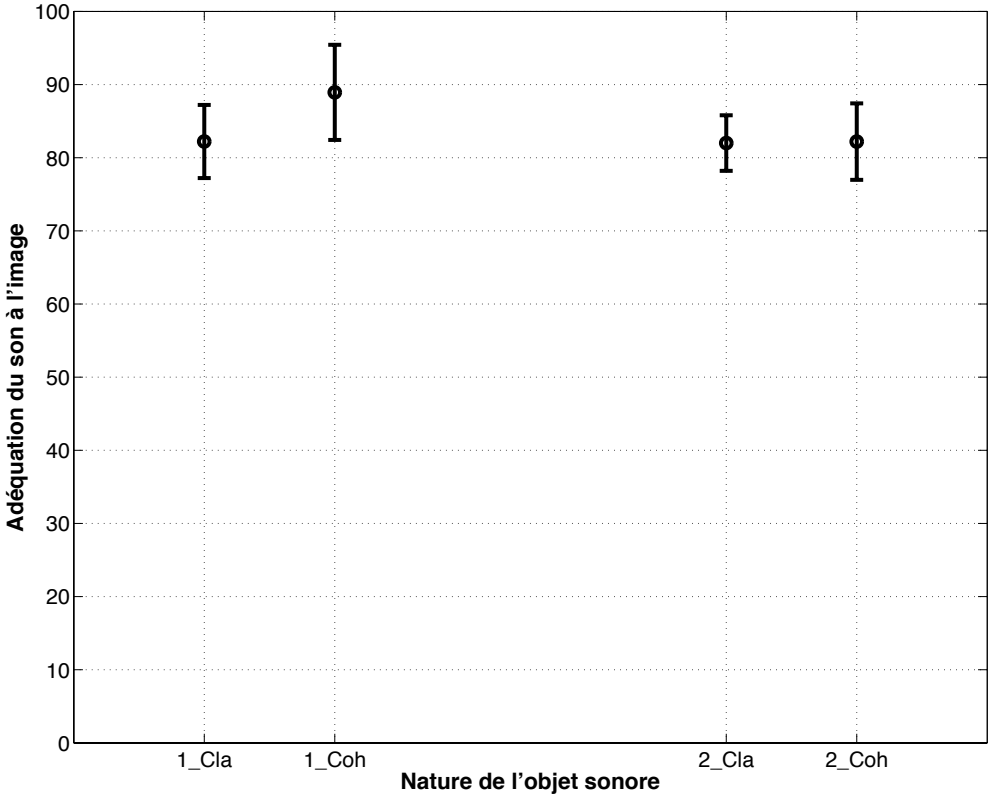


FIGURE 3.6 – *Effet de la cohérence azimutale sur l'adéquation du son à l'image en fonction de la nature de l'objet sonore. Notes moyennes et intervalles de confiance à 95%. 1 = effets sonores, 2 = voix, Cla = Mixage "classique", Coh = Mixage "cohérent"*

À ce stade on ne peut que faire un jugement global concernant la cohérence azimutale. Mais on peut d'ores et déjà affirmer que la cohérence azimutale n'est pas ressentie de la même manière quand l'objet sonore est une voix. L'interaction Az*N*V est également significative, il serait donc intéressant d'aller voir ce qu'il se passe plus en détail, en fonction de la vitesse des objets.

3.4.2 Corrélation entre la vitesse et la nature du stimuli

La significativité de l'interaction Az*N*V implique que la différence de jugement entre les scènes avec ou sans voix est également dépendant de la vitesse il apparait donc intéressant de regarder plus en détail la corrélation entre la vitesse et la nature de l'objet sonore.

Effets sonores

La figure 3.7 présente les évolutions d'adéquation du son à l'image en fonction des différentes vitesses proposées (les quatre états de vitesses présentés sont définis en 3.3.3).

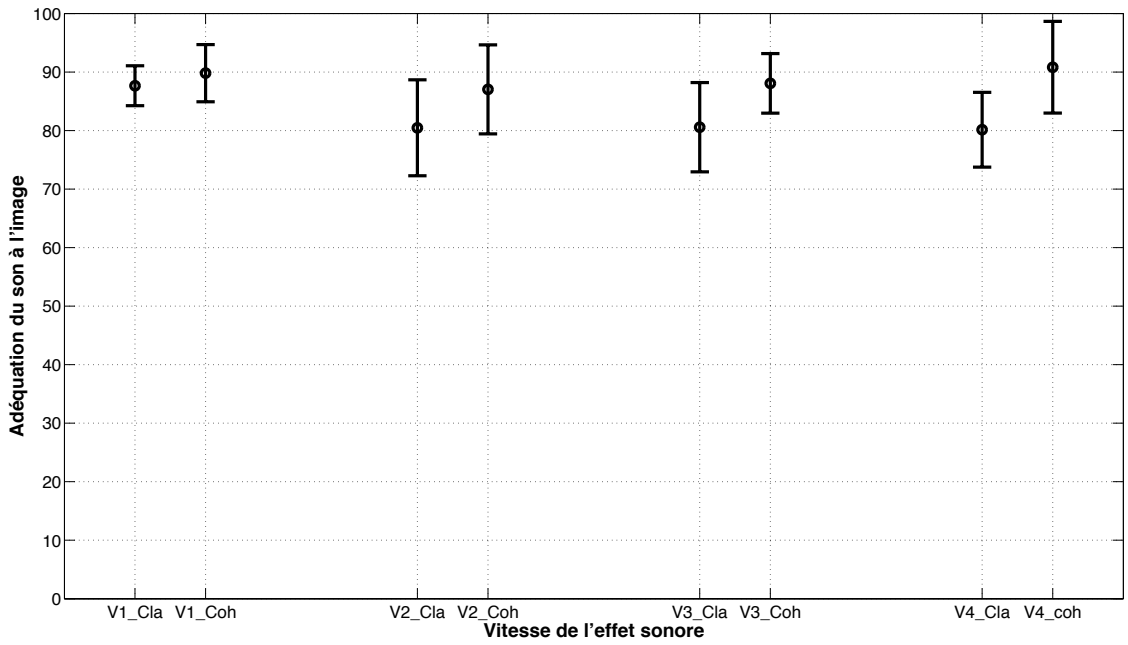


FIGURE 3.7 – *Effet de la cohérence azimutale pour les séquences avec effets sonores. Notes moyennes et intervalles de confiance à 95%. Cla = Mixage "classique", Coh = Mixage "cohérent", V*

Pour les séquences fixes, les sujets ne semblent pas avoir fait de réelle différence de jugement entre les versions "cohérentes" et "classiques". Bien qu'il y ai une moyenne plus élevée pour le mixage cohérent, la différence entre les deux notes n'est pas suffisamment significative.

Les résultats montrent en revanche que pour les états de vitesse 2 et 3 il y a une réelle amélioration sur le jugement d'adéquation. Bien que l'ANOVA ne considère pas ces différences comme significatives, on peut affirmer qu'il y a une tendance à préférer les mixages cohérents en azimut. Et il est fort à parier que ces résultats auraient été confirmés avec un plus grand nombre de sujets. Cependant, on ne semble pas observer d'augmentation réelle de l'amélioration avec l'augmentation de la vitesse de l'objet, comme c'était le cas sur l'expérience de Hendrickx.

Enfin, l'amélioration notée sur les versions avec montage est significative ($p = 0.014 \leq 0.05$). On peut donc affirmer que pour des effets sonores, les sujets ont préféré les mixages cohérents lorsqu'il y avait un montage image en jeu. Nous avions à priori imaginé que ce cas poserait problème notamment pour la séquence 1, il sera intéressant de voir les notes obtenues pour chaque séquence.

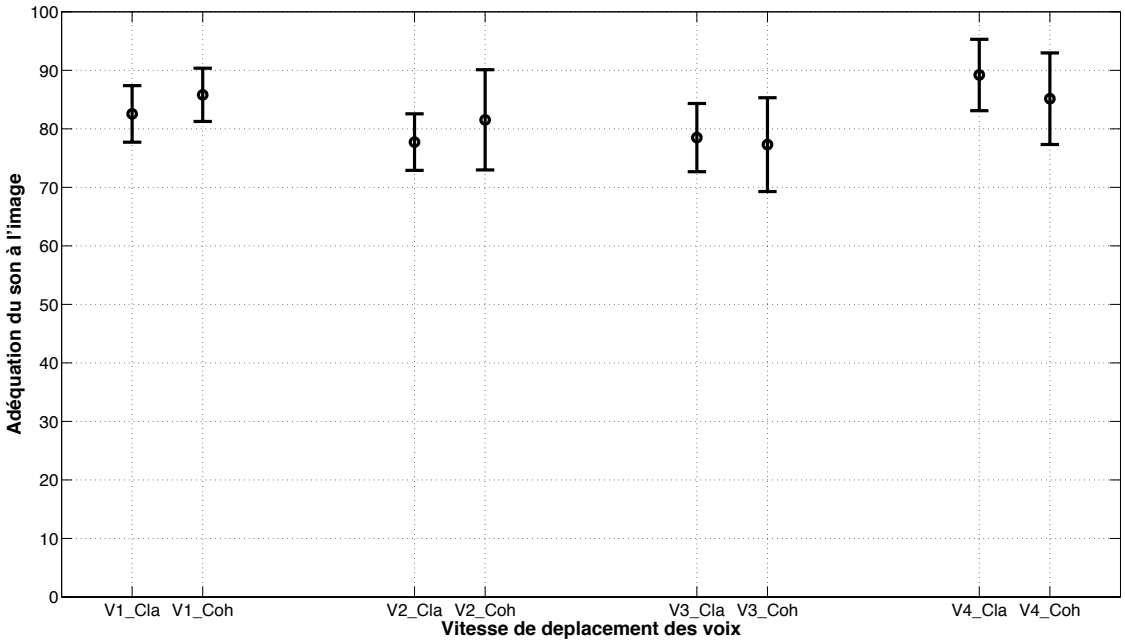


FIGURE 3.8 – *Effet de la cohérence azimutale pour les séquences avec voix. Notes moyennes et intervalles de confiance à 95%. Cla = Mixage "classique", Coh = Mixage "cohérent"*

Voix

La figure 3.8 présente les évolutions d'adéquation du son à l'image en fonction des différentes vitesses proposées.

Ici, il semble que pour les versions fixes et en mouvement lent les notes soient meilleures lorsqu'il y a une cohérence azimutale. Cependant, la différence de moyenne n'est pas suffisamment importante pour pouvoir l'affirmer avec autorité.

De plus, pour la voix, il semble n'y avoir aucune différence pour les mouvements rapides. En revanche, il est vraiment intéressant de noter que contrairement aux effets sonores, la hiérarchie semble s'inverser pour le montage image. Ici également, le résultat n'est pas significatif, mais il faudrait aller chercher plus en profondeur dans les différentes séquences, afin de voir si la non significativité de ce résultat ne provient pas uniquement des résultats d'une séquence en particulier. Effectivement, une des hypothèses avancées avant l'expérience, consistait à dire que le montage serait un point critique pour les séquences avec de la voix.

3.4.3 Influence de la vitesse de déplacement

Globalement la vitesse des objets ne semble pas être corrélées à l'amélioration ressentie par les sujets. Cette hypothèse a été confirmée par une analyse de l'amélioration notée par les sujets par rapport aux azimuts moyens, et maximums, et aux vitesses moyennes et maximales

des objets sonores à l'image. Nous allons cependant regarder les résultats pour chaque séquence en fonction de la vitesse, afin de voir si certaines de nos hypothèses se vérifient, et pour essayer de comprendre l'absence de résultats sur certains aspects de l'étude.

Séquence 1

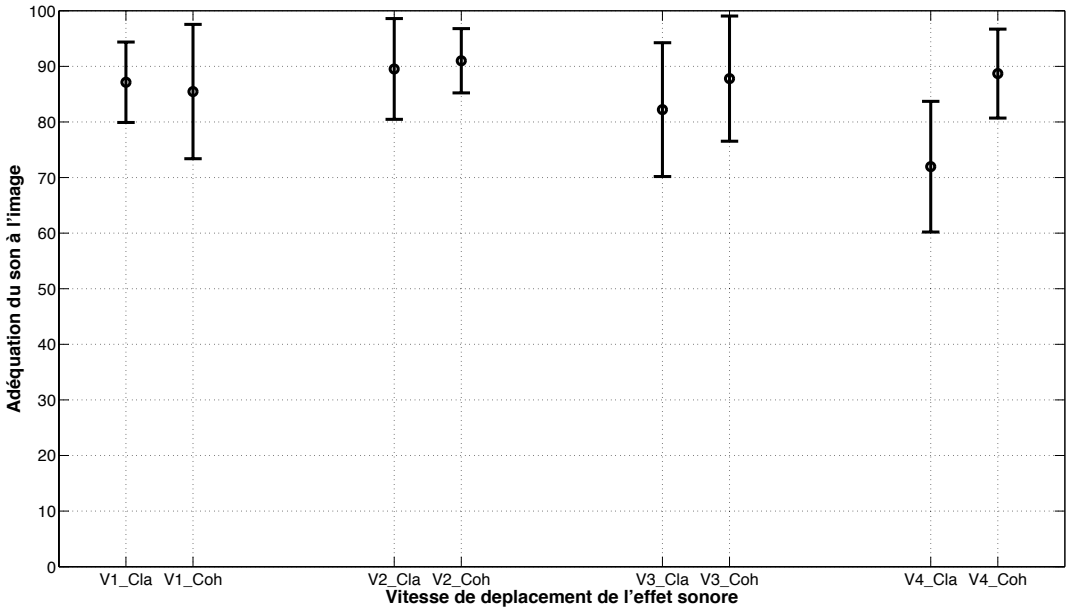


FIGURE 3.9 – *Effet de la cohérence azimutale pour la séquence 1 (roller). Notes moyennes et intervalles de confiance à 95%. Cla = Mixage "classique", Coh = Mixage "cohérent"*

Pour la séquence 1, il semble que les versions fixes et en mouvement lent n'ont pas provoqué de différences dans la manière de juger l'adéquation du son à l'image. Pour la vitesse plus élevée de l'objet, la différence semble déjà plus marquée entre le mixage cohérent et le mixage classique. Cependant, cela reste assez léger, et statistiquement non significatif. Mais au vu de la vitesse très importante de l'objet et la présence d'entrée et de sortie de champ très fréquente (cf tableau 3.3) on aurait pu attendre des différences plus importantes dans les résultats.

En revanche, pour la version comportant un montage, la version cohérente a été beaucoup mieux perçue que la version mixée sur l'enceinte centrale. Cette différence est significative avec $Sig.p = 0.023 \leq 0.05$. Cela veut dire que la version cohérente en azimut a réellement créé une différence quant à l'adéquation du son à l'image.

Ce résultat est surprenant puisque nous avions évoqué dans nos hypothèses que cette séquence poserait probablement problème du fait de la saute d'azimut importante avec une

énergie importante. Les résultats semblent donc contredire cette hypothèse. On peut imaginer que pour cette séquence très dynamique visuellement, la cohérence azimutale a pu augmenter cette sensation de mouvement et de dynamisme de la scène.

Version de la séquence	Vitesse Moyenne	Vitesse Maximale	Azimut Moyen	Azimut Maximal
Fixe	0˚/s	0˚/s	13.5˚	13.5˚
Mouvement lent	7.5˚/s	15˚/s	7.5˚	18.3˚
Mouvement rapide	23.9˚/s	90˚/s	16.8˚	22.5 ˚

TABLE 3.3 – *Azimut moyen, maximal et Vitesse moyenne et maximale de l'objet "roller" pour la séquence 1*

Coupe 1	Coupe 2	Coupe 3
-2˚-> -15˚	4.5˚-> -21˚	-19˚-> -13˚

TABLE 3.4 – *Saut d'azimut pour la version montée de la séquence 1*

Séquence 2

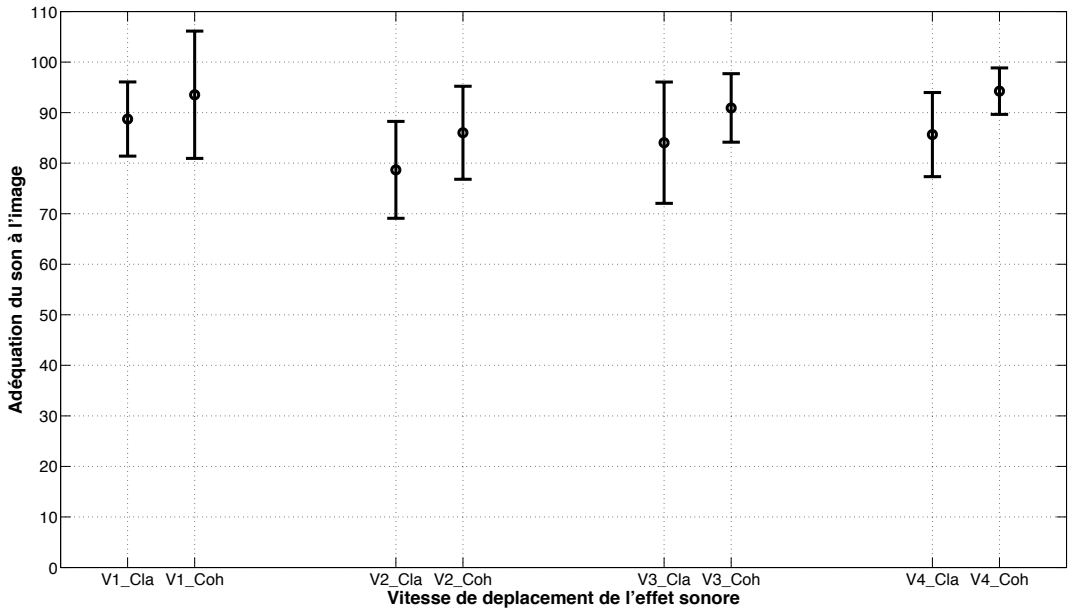


FIGURE 3.10 – *Effet de la cohérence azimutale pour la séquence 2 (mobylette). Notes moyennes et intervalles de confiance à 95%. Cla = Mixage "classique", Coh = Mixage "cohérent"*

Pour la séquence 2, il ne semble pas y avoir eu de réel impact de la vitesse de la mobylette, ni de l'azimut moyen. Si les notes sont globalement meilleures lorsque la séquence est mixée de manière cohérente, aucune tendance ne semble se dégager quant à l'influence des

vitesses. Ce résultat est assez surprenant, puisque sur les versions en mouvement et avec montage on avait systématiquement une sortie de champ, et une entrée de champ pour deux d'entre elle. Or, cela donne donc un effet très marqué, et des résultats plus significatifs avaient été envisagés pour le mouvement rapide.

Pour cette séquence également, la version montée a obtenu une différence significative ($p = 0.018 \leq 0.05$). Donc à nouveau, nous vérifions que la saute d'azimut ne semble pas poser problème aux sujets, et semble même être appréciée par ceux-ci.

Version de la séquence	Vitesse Moyenne	Vitesse Maximale	Azimut Moyen	Azimut Maximal
Fixe	0˚/s	0˚/s	8.5˚	8.5˚
Mouvement lent	2.6˚/s	23.2˚/s	4˚	22.5˚
Mouvement rapide	12.9˚/s	22˚/s	12.8˚	22.5 ˚

TABLE 3.5 – *Azimut moyen, maximal et Vitesse moyenne et maximale de l'objet "mobylette" pour la séquence 2*

Coupe 1	Coupe 2
8.5˚-> 10˚	-7˚-> 22.5˚

TABLE 3.6 – *Saut d'azimut pour la version montée de la séquence 2*

Séquence 3

Pour la séquence 3, il semble qu'il y ait eu un impact global de la présence de mouvement, mais sans différence en fonction de la vitesse de celui-ci. Cependant, on peut noter à nouveau que ces résultats ne sont pas statistiquement significatifs.

Comme sur les séquences 1 et 2, la version montée présente à nouveau une différence importante. Ici l'indice de significativité vaut $Sig.p = 0.078$ ce qui indique que cette différence serait peut-être devenue significative avec plus de sujets.

Version de la séquence	Vitesse Moyenne	Vitesse Maximale	Azimut Moyen	Azimut Maximal
Fixe	0˚/s	0˚/s	11.5˚	11.5˚
Mouvement lent	2.6˚/s	4.2˚/s	12.8˚	22.5˚
Mouvement rapide	4.8˚/s	13.9˚/s	12˚	20.2 ˚

TABLE 3.7 – *Azimut moyen, maximal et Vitesse moyenne et maximale de l'objet "marteau" pour la séquence 3*

Coupe 1	Coupe 2	Coupe 3
-14˚-> -22.5˚	-22.5˚-> 5˚	-8˚-> -12˚

TABLE 3.8 – *Saut d'azimut pour la version montée de la séquence 3*

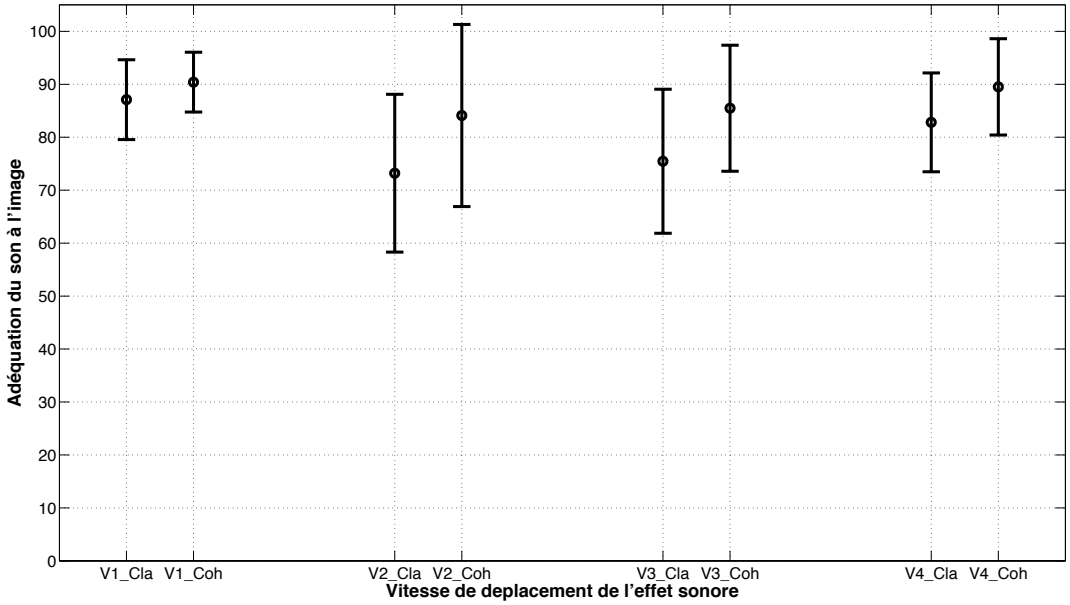


FIGURE 3.11 – *Effet de la cohérence azimutale pour la séquence 3 (marteau). Notes moyennes et intervalles de confiance à 95%. Cla = Mixage "classique", Coh = Mixage "cohérent"*

Séquence 4

Pour la séquence 4, il semble clair que les sujets n'ont pas fait de différences entre les mixages cohérents et classiques. Cela n'implique pas que la cohérence azimutale n'a pas été détectée, mais que ce paramètre n'a pas eu d'importance dans le jugement fait par les sujets.

Il est intéressant de remarquer qu'en terme de cadre et d'action, la séquence 4 est très proche de la séquence 2 mettant en scène une mobylette. Effectivement, la source est en mouvement, avec des entrées et sorties de champs, et le cadre est très large. Cependant, les résultats ne sont clairement pas les mêmes. Si dans les deux séquences il ne semble pas y avoir d'impact de la vitesse de l'objet, les sujets avaient jugé globalement les mixages cohérents avec des meilleures notes pour la mobylette, alors que pour la voix des joggeuses cela n'a pas du tout influé sur la note. Cela confirmerait que les sujets n'ont pas adopté la même stratégie selon la nature de l'objet sonore.

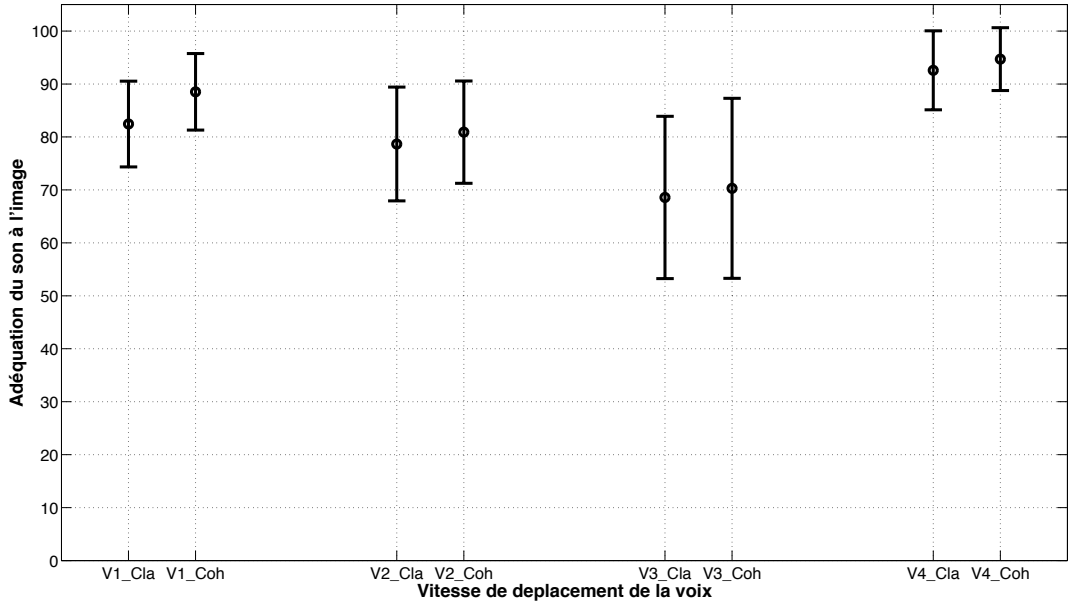


FIGURE 3.12 – *Effet de la cohérence azimutale pour la séquence 4 (Voix des joggeuses). Notes moyennes et intervalles de confiance à 95%. Cla = Mixage "classique", Coh = Mixage "cohérent"*

Version de la séquence	Vitesse Moyenne	Vitesse Maximale	Azimut Moyen	Azimut Maximal
Fixe	0˚/s	0˚/s	10.5˚	10.5˚
Mouvement lent	3˚/s	9.2˚/s	10.3˚	22.5˚
Mouvement rapide	5˚/s	6˚/s	14.9˚	22.5 ˚

TABLE 3.9 – *Azimut moyen, maximal et Vitesse moyenne et maximale de l'objet "voix joggeuses" pour la séquence 4*

Coupe 1	Coupe 2
22.5˚-> -22.5˚	22.5˚-> -22.5˚

TABLE 3.10 – *Saut d'azimut pour la version montée de la séquence 4*

Séquence 5

Pour la séquence 5, on retrouve la tendance que l'on avait sur la figure 3.8, il semble que pour la version fixe et en mouvement lent, le mixage cohérent a été jugé plus adapté à l'image. Cependant, le manque de sujets rend ces écarts non significatif, et ce malgré une différence de moyenne de plus de 10 points pour le mouvement lent. Il faudrait vérifier ce résultat avec un plus grand nombre de sujets, pour voir si la tendance se confirme.

Pour la version avec mouvement rapide et la version montée, à nouveau, les sujets n'ont pas fait de différence dans leur façon de juger l'adéquation du son à l'image, en fonction de la cohérence azimutale.

Version de la séquence	Vitesse Moyenne	Vitesse Maximale	Azimut Moyen	Azimut Maximal
Fixe	0˚/s	0˚/s	17˚	17˚
Mouvement lent	4.2˚/s	8.2˚/s	10.1˚	17.5˚
Mouvement rapide	6.9˚/s	16.4˚/s	8.1˚	15.9 ˚

TABLE 3.11 – *Azimut moyen, maximal et Vitesse moyenne et maximale de l'objet "voix d'homme" pour la séquence 5*

Coupe 1	Coupe 2	Coupe 3
17˚-> -15˚	-15˚-> -3˚	-1.5˚-> -8˚

TABLE 3.12 – *Saut d'azimut pour la version montée de la séquence 5*

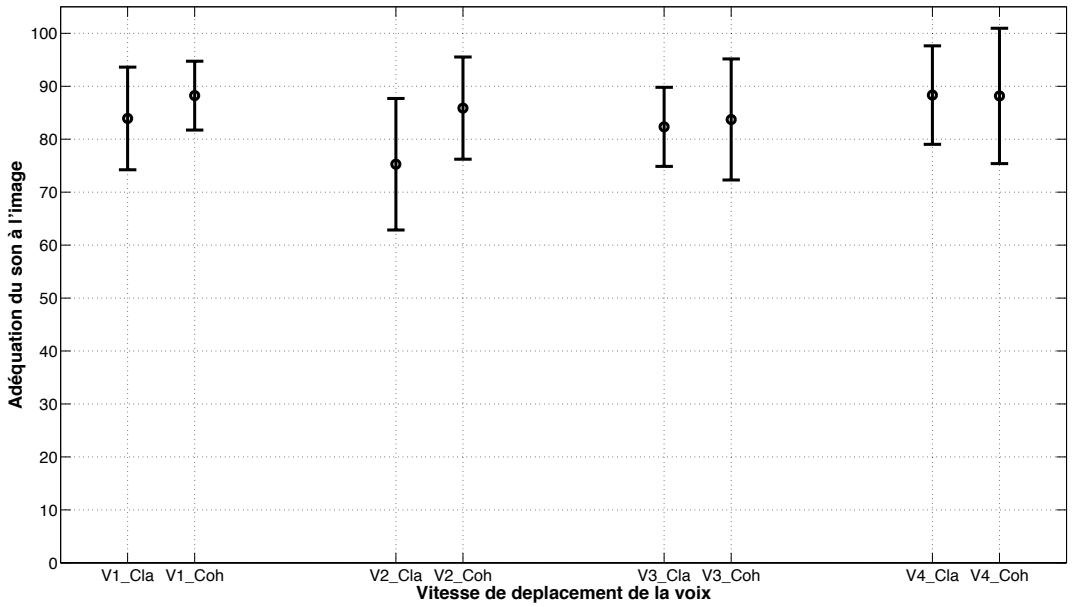


FIGURE 3.13 – *Effet de la cohérence azimutale pour la séquence 5 (voix de l'homme). Notes moyennes et intervalles de confiance à 95%. Cla = Mixage "classique", Coh = Mixage "cohérent"*

Séquence 6

Pour la séquence 6, on remarque que comme pour la séquence 4, il n'y a pas eu d'impact de la cohérence azimutale pour les versions fixe et en mouvement lent. Cependant, il est très intéressant de remarquer que pour la vitesse rapide les tendances semblent s'inverser, et la

version cohérente est jugée moins bien que la version classique. De plus, cette tendance est encore plus marquée pour la version avec montage image.

Nous avions évoqué avant d'analyser les résultats du test que cette séquence poserait très certainement problème, notamment pour la version avec montage, puisqu'elle présente un champ-contrechamp. Ces résultats confirment donc que la cohérence azimutale sur le cas du champ-contrechamp est globalement mal perçue par un sujet "lambda". Cependant, bien qu'il y ait une différence de moyenne de 17 points, le nombre insuffisant de sujets rend cet écart non significatif, avec un indice de significativité de 0.081. On peut cependant aisément affirmer que ce résultat aurait été significatif avec plus de sujets.

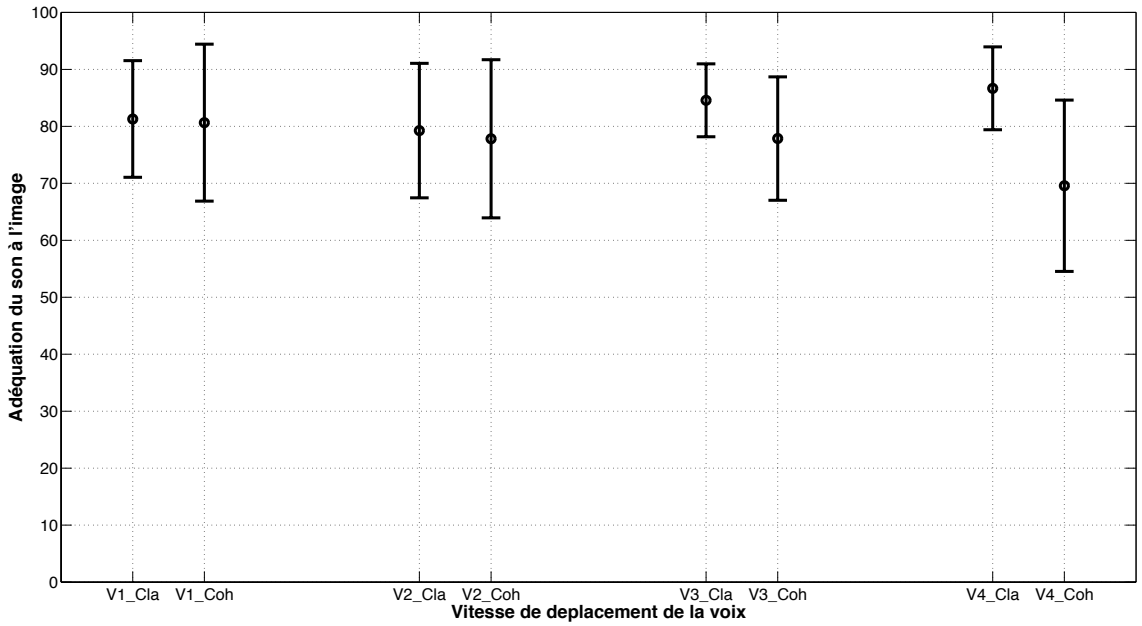


FIGURE 3.14 – *Effet de la cohérence azimutale pour la séquence 6 (Discussion père-fille). Notes moyennes et intervalles de confiance à 95%. Cla = Mixage "classique", Coh = Mixage "cohérent"*

Version de la séquence	Vitesse Moyenne	Vitesse Maximale	Azimut Moyen	Azimut Maximal	Azimut de la fille
Fixe	0˚/s	0˚/s	12.5˚	12.5˚	-9.5˚
Mouvement lent	4.3˚/s	9.4˚/s	8.4˚	15.15˚	0˚
Mouvement rapide	5.3˚/s	18.4˚/s	6.7˚	19.77 ˚	-12˚

TABLE 3.13 – *Azimut moyen, maximal et Vitesse moyenne et maximale des objet "Voix du père" et "Voix de la fille"pour la séquence 6*

Coupe 1	Coupe 2	Coupe 3	Coupe 4	Coupe 5	Coupe 6	Azimut de la fille
-12.5˚-> -22.5˚	-22.5˚-> 9.5˚	-9.5˚-> -22.5˚	-22.5˚-> -7˚	-7˚-> -22.5˚	-11.5˚-> 0 ˚	9.5˚

TABLE 3.14 – *Saut d'azimut pour la version montée de la séquence 6*

3.4.4 Discussion avec les sujets

Chaque sujet a rempli un questionnaire post-test afin de recueillir leur impression sur le test, ainsi que pour voir si ils avaient détectés les paramètres variant. C'était également l'occasion de discuter avec eux.

Tout d'abord, de manière assez surprenante de nombreux sujets n'ont pas détecté que les sons étaient latéralisés ou non. Un sujet a même avoué ne pas avoir détecté de différence de mixage durant son test[5]. Et si plusieurs ont détecté la cohérence azimutale, uniquement deux sujets sur treize ont supposé, à raison, que c'était le seul paramètre qui variait. L'un d'entre eux m'a également confié qu'il avait trouvé que la latéralisation n'était selon lui pas tout le temps nécessaire et parfois trop exagéré. De même, un sujet m'a expliqué qu'il trouvait que l'image amenait déjà la sensation de mouvement, et que la cohérence azimutale du son n'était pas forcément nécessaire.

De plus, de nombreux sujets ont eu des sensations de changement sur des paramètres qui n'avaient pourtant pas changé. Aussi, de nombreuses personnes ont eu la sensation de modifications du volume de certaines sources sonores, et parfois d'évolution du niveau durant la séquence, alors même qu'il n'y avait qu'un mouvement en azimut du son, et sachant que les enceintes avaient été réglées afin d'avoir un niveau équivalent sur les deux versions du mixage.

De même, plusieurs personnes ont ressenti des variations quant aux réverbérations dans les différents lieux. Enfin, plusieurs sujets ont supposé des modifications sur le niveaux des ambiances et la présence ou non de certains sons d'ambiances. Ces deux remarques sont assez intéressantes, puisque l'on se rend compte que le fait de faire suivre le son de manière cohérente semble avoir démasqué certaines sources.

5. Ce sujet n'a pas été écarté, car ses résultats ne traduisaient pas une mauvaise compréhension de la question comme le sujet 5.

3.5 Conclusion

La réalisation de cette expérience a soulevé plusieurs intérêts, ainsi que certaines limites du test en lui même.

Tout d'abord, nous avons pu en partie vérifier certaines de nos hypothèses. En effet, comme nous l'avions prévu, les sujets n'ont pas jugé de la même manière les séquences avec de la voix et celle avec des objets classiques. De plus, le champ contre-champ semble être comme attendu un cas limite pour la cohérence azimutale.

Si nous vérifions effectivement que la cohérence azimutale apporte une plus grande amélioration quand il y a un mouvement, nous n'avons pas pu retrouver une corrélation entre la vitesse du mouvement et l'amélioration apportée.

Nous avons pu voir que contrairement à nos hypothèses, les versions avec montages pour les effets sonores ont largement été préférées lorsque le mixage était "cohérent", et ce y compris pour la séquence 1 que nous avions jugé à priori plus gênante en mixage "cohérent".

Enfin, il est important de relever que si, de manière globale, les sujets n'ont pas fait de différences selon les mixages pour les séquences avec voix, ils n'ont donc pas non plus juger plus sévèrement les versions "cohérentes" (hormis pour le champ contre-champ de la séquence 6). Cela pourrait montrer que l'effet ventriloque a été assez fort sur les séquences avec de la voix. Dans ce test, nous avions une différence maximale de 22.5˚, or plusieurs études ont démontré un effet ventriloque pour des disparités allant jusqu'à 20˚(Thurlow et Jack 1973, Komiyama 1989) et même jusqu'à 30˚(Jackson, 1953).

Nous avons néanmoins pu voir les limites de ce test, puisque le manque de sujets a rendu assez compliqué l'exploitation des résultats. En effet, de nombreuses différences de résultats semblant pourtant importantes, ont été jugées statistiquement non-significatives, rendant donc plusieurs conclusions difficiles à affirmer avec autorité. Il serait donc intéressant de voir les résultats que l'on obtiendrait avec ce même test, pour un nombre de sujets plus important. Cela confirmerait peut-être les tendances que l'on a pu observer dans les résultats.

Conclusion générale

Le but de ce mémoire était de questionner l'apport de la cohérence audiovisuelle spatiale en azimut dans le cadre d'une exploitation cinématographique. L'étude historique de la spatialisation du son au cinéma nous a permis de voir, que pendant très longtemps ce sont avant tout des contingences techniques et économiques qui ont dicté notre manière d'aborder la spatialisation des sons. Ce n'est que relativement récemment, depuis les années 1990, que les technologies sont devenues suffisamment précises pour donner enfin tout droit aux réalisateurs et aux mixeurs sur la spatialisation des sources sonores. Les nouveaux systèmes semblent de plus permettre une meilleure cohérence et des outils de spatialisation plus performants.

Nous pouvons cependant noter que la cohérence audiovisuelle spatiale est forcément à mettre en regard avec le questionnement autour de la fonction du cinéma qui se pose depuis ses débuts. Celui-ci doit-il être une simple trace du réel et donc s'attacher à le représenter de la manière la plus cohérente possible ? Il semble que le cinéma soit plus que ça. En tant qu'art, il ne fait pas que représenter le réel, mais il le transcende, en fait quelque chose d'autres. Il est donc légitime de se demander si la cohérence audiovisuelle spatiale a réellement sa place dans ce contexte, et surtout pour la voix.

Si la cohérence azimutale est très fréquente et plutôt appréciée pour des sources sonores non réalistes, ou plus proches du bruitage, les cas où la voix n'est pas mixée au centre semble être assez particulier. Aussi, ce sont plutôt les voix au statut ambigu qui se voient adjuger un traitement différent quant à leur placement : les voix intérieures, les voix sans corps, les voix entre deux mondes (à la fois dans et hors de la diégèse). On recense peu de cas de cohérence azimutale pour la voix, et quand c'est le cas (*Gravity*, *Strange Days* ou *Birdman*), l'effet fait partie intégrante de l'écriture du film.

Afin de comprendre pourquoi il était acceptable pour un spectateur de percevoir un stimulus audiovisuel cohérent temporellement mais pas spatialement, nous avons tâché d'étudier les mécanismes reliant la perception du son et de l'image. Nous avons pu voir que la perception de l'espace pour l'ouïe était fortement influencée par les informations reçues par la perception visuelle. Aussi, des effets de fusion entre deux événements spatialement séparés se manifestent jusque 30˚de séparation, c'est l'effet ventriloque. Nous avons également émis l'hypothèse que

moins le spectateur a une approche discriminante d'un stimulus audiovisuel, plus l'effet ventriloque est efficace, ce qui tend à confirmer que l'effet ventriloque serait suffisant pour faire correspondre la voix et son corps associé au cinéma.

En nous basant sur différentes études ayant montré que la cohérence audiovisuelle spatiale pouvait avoir un impact positif sur l'adéquation du son à l'image, nous avons mis au point un test perceptif visant à décortiquer les critères impactant le jugement d'adéquation du son à l'image en terme de spatialisation. Treize sujets ont vu 24 séquences mettant en scène 6 sources sonores différentes, trois comportaient des voix et les trois autres des effets sonores (une mobylette, des rollers et un marteau). Pour chaque source sonore quatre situations étaient utilisées :

- 1 : La source sonore est fixe

- 2 : La source sonore est en mouvement lent

- 3 : La source sonore est en mouvement rapide

- 4 : La source sonore se déplace "instantanément" par le biais d'un découpage de la séquence.

Afin de tester l'impact de la cohérence azimutale, chaque séquence était présentée dans un ordre aléatoire avec un mixage classique où le son était diffusé au centre et un mixage cohérent où la source sonore suivait son correspondant visuel à l'image. Les sujets devaient juger pour chaque séquence à quel point le son était adapté à l'image.

L'analyse des résultats nous a permis de confirmer certaines des hypothèses de ce mémoire. Comme attendu, les séquences avec de la voix placée de manière cohérente ont été jugées différemment que les autres. Si pour la mobylette, les rollers et le marteau, la cohérence semble avoir amélioré l'adéquation du son à l'image, pour la voix de manière globale, la cohérence azimutale n'a eu aucun effet sur le jugement. En regardant dans les détails, on se rend compte que la cohérence en azimut a même parfois été jugée négativement, c'est le cas de la scène utilisant un découpage de type champ contre-champ. Ces résultats tendraient donc à confirmer que la cohérence audiovisuelle spatiale pour les voix ne semble pas être importante pour l'appréciation d'une séquence par les spectateurs, il semblerait même que cela puisse gêner ces derniers lorsque l'on est en présence d'un montage image. Or, la majorité des créations cinématographiques actuelles sont extrêmement découpées.

Le manque de sujets n'a malheureusement pas permis de relever d'autres conclusions statistiquement pertinentes. Il serait donc intéressant de renouveler l'expérience avec un nombre plus important de sujets afin de voir si les tendances au sein des résultats se précisent.

Il faut cependant souligner que, dans le cadre de ce test perceptif, la cohérence audiovisuelle spatiale a été appliquée scrupuleusement, sans aucun soucis d'une réelle écriture sonore, ni sans ajustement de placement par le mixeur autre que la position exacte des sources à l'écran. Plusieurs exemples de films cités dans la première partie nous montrent que la cohérence audiovisuelle spatiale en azimut pour les voix peut être un vrai apport pour un film si ce dernier s'y prête.

Ce mémoire n'a donc pas pour finalité de condamner la cohérence azimutale pour les voix au cinéma, mais montre simplement, en l'état, que l'application pure et bête de la cohérence en azimut ne semble pas améliorer l'expérience audiovisuelle des spectateurs, dès lors qu'elle ne découle pas d'une écriture sonore particulière au film.

Annexes

ANNEXE 1 : Écriture des séquences

Dès que l'idée de réaliser un test perceptif m'est venu, j'ai statué sur le fait que les séquences employées devaient s'éloigner le plus possible de séquences "expérimentales" comme on en retrouve dans beaucoup de test perceptif (notamment sur l'effet ventriloque), avec par exemple un homme en pied lisant un texte (qui dans le cas de mon expérience pourrait se déplacer). En effet, les stimuli présentés sont souvent très éloignés du contexte réellement étudié, que ce soit une situation de la vie de tous les jours, où dans notre cas la projection d'un programme audiovisuel. C'est pourquoi j'ai fait le choix de montrer des séquences pouvant appartenir à des fictions cinématographiques.

Cependant, les conditions de mon test ne me permettent pas de réutiliser des séquences de films pré-existants, comme c'est le cas (en partie) dans l'expérience V de la thèse de Hendrickx, ou dans les tests du mémoire sur la WFS au cinéma de Carreau et Macquart. En effet, pour obtenir les différentes situations avec une seule source, il était nécessaire de me créer moi-même la matière visuelle et sonore. J'ai donc écrit six séquences se prêtant aisément à mon sujet et qui pourraient prendre place au sein d'un film de fiction.

Objets sonores

Les séquences ayant pour source sonore principale un effet sonore furent assez simple à concevoir. En effet, il est plus habituel au cinéma de faire se déplacer des bruitages ou des effets sonores dans un mixage.

J'ai donc d'abord décidé de mettre en scène une mobylette qui roulerait sur une route de campagne vide, aussi le son de la mobylette est à la fois continu, très reconnaissable et assez riche spectralement. De plus, le son de la mobylette évolue peu selon les régimes et on peut obtenir le même son et à l'arrêt et en roulant, ce qui correspond à ce que j'attendais pour mon test.

La seconde séquence met en scène un joueur de roller hockey s'entrainant sur un parking. L'intérêt du roller est qu'on peut obtenir des changements de rythme très intéressant pour mon sujet, et le son est un peu moins large que celui de la mobylette. Cependant, ce son est à nouveau un son continu, de plus, c'est également un son associé à un moyen de locomotion, ce qui est très souvent associé au cinéma avec des panoramiques d'effets sonores. Il me fallait donc trouver un autre type de source.

Dans ce but, la troisième séquence met en scène un homme clouant un tasseau sur une planche. Le son du marteau contre le clou est à la fois large spectralement, percussif et discontinu, ce qui me permet de tester mes hypothèses sur d'autres types d'événements.

Afin de créer les mouvements nécessaires au test, il a donc fallu utiliser des mouvements de caméra, la source sonore ne pouvant réellement se déplacer.

La voix

Il a été plus compliqué de préparer les séquences avec de la voix. En effet, les contraintes que je me suis moi-même imposé pour réaliser le test perceptif sont difficiles à prendre en compte pour concevoir une séquence, notamment par l'obligation d'avoir des mouvements.

La première de ces séquences, met en scène deux joggeuses dans un parc. Pour correspondre aux différentes situations, elles ont été filmées soit en courant, en marchant (après avoir couru), ou à l'arrêt avant de courir. Il a donc fallu trouver un texte assez neutre pour marcher dans les différents cas souhaités. Bien évidemment le jeu des actrices s'adapte aux différentes situations, donc la source sonore n'est pas entièrement la même selon les plans. Cependant, en ayant le même texte, le même cadre et les mêmes actrices, on limite les paramètres variant entre les différentes versions.

La deuxième séquence, présente un jeune homme seul qui sort dans son jardin pour téléphoner à sa mère. L'action de téléphoner est facilement rattachable à l'idée de mouvement, en effet nombreuses sont les personnes qui font les 100 pas en téléphonant.

Enfin, la dernière séquence permet de tester une situation où deux personnages sont physiquement éloignés l'un de l'autre à l'écran. On y montre un homme d'une quarantaine d'année qui dispute sa fille qui n'est pas rentrée dormir la veille. La fille est assise sur un canapé tandis que son père reste debout. L'intérêt de cette séquence est de pouvoir spatialiser deux voix à deux endroits différents dans la même séquence. De plus, la version découpée est conçue comme un champ-contrechamp. Or, c'est le poncif de ce qui peut poser problème quant à la spatialisation du son lors du mixage. C'est donc l'occasion de voir si l'effet de déplacement des voix en azimut posera un vrai problème dans cette situation classique de découpage d'une séquence.

ANNEXE 2 : Listes complètes du matériel de tournage

Matériel image

La caméra utilisée lors des tournages était une Blackmagic production camera 4K, couplée avec deux optiques à focale fixe, une optique Zeis MacroPlanar 50mm f2 et une Samyang 24mm f1.4. L'enregistrement se fait directement sur disque dur SSD (Sandisk Extreme pro 480 Go). La caméra est associée à une Matte box cinematics. Enfin, nous avons utilisé un pied manfrotto 055XB associé à une rotule vidéo 701HDV. Afin de pouvoir monitorer également la vidéo pendant les prises, un moniteur Blackmagic Video Assist était installé sur la caméra sur un petit bras magique.

Matériel son

Voici la liste du matériel son utilisé pour le tournage des séquences du test perceptif :
- Un enregistreur Sounddevices 664
- Deux ensembles HF émetteur et récepteur Audiolimited 2020
- Trois microphones cravate Sanken cos 11
- Un microphone cardioïde Senheiser MKH50
- Un microphone semi-canon Senheiser MKH416
- Un couple MS Schoeps composé de deux corps de micro CMC6 associés à une capsule MK4 et une capsule MK8
- Une suspension Schoeps AMSCI avec Bonnette WSRMSCI et Windjammer WJMS/XY
- Une suspension rycotte avec Bonnette et windjammer pour MKH416
- Une suspension cinela osix pour MKH50
- Une suspension cinela osix pour MKH416
- Une perche moyenne VDB (4m) et une perche courte VDB (2,5m)
- Un pied de micro et un pied de table K&M
- Une L48
- Trois batteries NP1 avec leur chargeur
- Un adaptateur NPC Hawkwood pour alimenter la 664 et les HF

ANNEXE 3 : Méthodologie de tournage

Séquence 1

FIGURE A.1 – *Capture d'écran de la version avec mouvement lent de la séquence 1*

Dans cette première séquence l'objet sonore choisi est le son des rollers sur le bitume du parking. L'environnement sonore était assez chargé, à savoir une départementale à une centaine de mètres. Il était donc indispensable d'avoir une prise de son au plus proche des rollers.

Un micro cravate a donc été placé directement sur le côté intérieur d'un des rollers avec l'émetteur HF dans la poche de l'acteur. Ce dernier ne capte donc que le son des rollers.

On retrouve néanmoins une perche avec MKH416 en liaison HF également, qui était parfois cachée derrière les voitures pour des plans très larges. Sur les plans en mouvement, le MKH50 sur pied est placé sur le côté opposé de là où se trouve le perchman. Les mouvements du personnage étant amples, le micro sur pied permet de compléter la prise de son pour couvrir tout son mouvement. Le couple MS Schoeps est quant à lui placé dans l'axe de la caméra.

Séquence 2

Dans cette seconde séquence l'objet sonore est le son du moteur de la mobylette. Comme sur la séquence 1, un micro cravate a été placé sur la mobylette près du moteur, avec le boitier HF fixé sur la carcasse de la mobylette de sorte qu'on ne le voit pas à l'image.

FIGURE A.2 – *Capture d'écran de la version fixe de la séquence 2*

Le son de la mobylette étant nettement plus fort que celui des rollers, le micro perche (MKH416) bien que placé assez loin de la mobylette, capte relativement bien le son de la mobylette. De plus les plans sont assez larges, et le son capté par la perche est donc assez adéquat. Cependant, le HF est nécessaire dans le but de pouvoir séparer le son de la mobylette de son environnement. Comme précédemment, un couple MS est placé dans l'axe de la caméra.

Séquence 3

FIGURE A.3 – *Capture d'écran de la version fixe de la séquence 3*

Dans cette troisième séquence l'objet sonore mis en scène est un bruit de marteau contre un clou. Ici le son du marteau est très fort, je n'ai donc pas placé de micro cravate, le son capté par le micro perche étant suffisant. Ici, le micro principal est un MKH50. Un couple MS a également été placé dans l'axe de la caméra. Cependant, la scène a dû être tournée dans un garage dont la porte était ouverte, alors même que l'on ne voyait que l'intérieur du garage, de ce fait l'ambiance captée par le couple la plupart du temps me semble trop capter l'environnement sonore de la ville et donc trop peu adapté à la scène tournée (cela reste cependant à vérifier).

Séquence 4

La séquence 4 est la première scène avec du dialogue. Ici l'objet sonore est le groupe composé par les deux femmes qui font leur jogging.

FIGURE A.4 – *Capture d'écran de la version avec mouvement rapide de la séquence 4*

Pour obtenir l'effet recherché, j'ai dû tourner des plans très larges et donc très difficiles à percher. Chaque actrice portait donc un micro cravate, dont le placement était crucial, puisque la majorité du dialogue est exclusivement capté par ces micros. La perche avec un MKH416 capte le dialogue un peu plus lointain et les pas des joggeuses. Il faudra cependant voir si il est réellement possible de se servir de la perche dans une optique de spatialisation de la voix, sachant qu'elle capte énormément de l'environnement sonore attenant. Le couple MS est placé dans l'axe de la caméra, et il capte à la fois l'ambiance du parc qui est assez calme avec du vent dans les arbres et des oiseaux, mais également le champ diffus de la réverbération induite par les arbres.

Cependant, les actrices courent sur plusieurs plans et le son capté par les HF n'est pas toujours très propre. Le texte a donc été réenregistré en son seul à l'arrêt, où la perche

pouvait également capter le dialogue. Cela offrira plus de souplesse lors du montage son de la séquence.

Séquence 5

FIGURE A.5 – *Capture d'écran de la version avec mouvement lent de la séquence 5*

Cette cinquième séquence met en scène un homme au téléphone, l'objet sonore est donc la voix de ce personnage. Dans la scène il se tient dans son jardin devant sa véranda ouverte, et il s'adresse à sa femme par la véranda depuis l'extérieur.

Le cadre étant assez bas, la perche capte très bien le dialogue. Pour la compléter un micro cravate est placé dans le col de la chemise. Quand l'homme s'adresse à sa femme la perche étant à l'extérieur ne peut pas parfaitement timbrer la voix, un MKH50 a donc été caché dans la véranda afin de capter les répliques s'adressant à la femme.

Séquence 6

Dans la dernière scène on a deux personnages qui se parlent mais qui contrairement à la séquence 4, ne sont pas au même endroit. De plus, l'un est debout, tandis que l'autre est assis. Or, il est crucial pour le test que les deux voix soient bien séparées. Hormis sur le champ-contrechamp, les deux personnages ne sont pas du tout à la même hauteur dans l'image, et donc le micro perche (MKH416) ne peut nécessairement pas capter avec la même présence la voix du père et celle de sa fille. Un MKH50 a donc été caché dans le décor sur chacun des plans pour capter la réplique de la fille assise. Ce même micro a capté le OFF sur le champ-contrechamp.

En complément, chaque personnage porte un micro cravate. Enfin, un couple MS permet de capter un peu de champ réverbéré de la salle, et les mouvements de l'acteur.

Sur chaque décors, des ambiances ont été enregistrées ainsi que des silences raccords dans les lieux plus calmes.

FIGURE A.6 – *Capture d'écran de la version fixe de la séquence 6*

ANNEXE 4 : Détails sur la post-production des séquences

Le montage des ambiances

Pour se rapprocher le plus possible de séquences de cinéma classique, un montage d'ambiances et d'effets a été réalisé en 5.0. Les sons constituant ces ambiances provenaient à parts égales de sonothèque et d'enregistrements effectués directement par l'auteur, sur les lieux de tournages des séquences, ainsi que sur des tournages précédents. Ces enregistrements ont été effectués pour la grande majorité en MS ou en DoubleMS.

Lors du montage son, je vérifiais régulièrement le rendu lorsque la source était spatialisée. Bien que la spatialisation à ce stade soit assez grossière et uniquement en LCR, cela me permettait de vérifier d'éventuels problèmes de déséquilibre de la scène sonore. Effectivement, si l'objet sonore principale de la scène se déplace ou est fixé à un endroit excentré, le risque est de se retrouver avec une espace sonore déséquilibré. Le montage son s'est donc appliqué à réduire ces déséquilibres.

De même, un des enjeux était également de nourrir en permanence le canal central, pour que le mixage ne soit pas uniquement des ambiances en stéréo et un objet sonore se déplaçant entre les deux canaux gauche et droite. Dans toutes les scènes on retrouve des ambiances monophoniques, parfois issues des ambiances MS, dans le canal de centre, afin d'avoir une assise dans le mixage de la séquence.

Pour limiter les paramètres variant entre les différentes versions des séquences, le montage son est quasiment identique entre ces versions, hormis une adaptation à la valeur de plan et à la durée du plan. Aussi, le seul paramètre variant sera la vitesse de déplacement de l'objet sonore et non l'environnement sonore attenant.

Enfin, il s'est présenté pour deux séquences [6] la question du placement des bruitages. En effet, aujourd'hui la plupart des mouvements sont renforcés par le bruitages et tous les pas sont refaits par le bruiteur, soit pour les appuyer, soit pour préparer la VI (Version Internationale). Or ces deux séquences méritaient un renfort sur les bruits de pas qui n'étaient pas assez nets. Des pas de sonothèques ont donc été synchronisés avec l'image. Il m'a paru logique de latéraliser les bruits de pas et la voix de la même manière, puisque c'est le même corps qui est à l'origine des deux sons. Si cela semble assez évident, il est cependant nécessaire de se poser la question de la place des bruitages dans un mixage où la voix serait déplacée du centre. Cela ne doit pas se faire au détriment du couple voix-bruitages.

6. La séquence 4 mettant en scène des joggeuses et la séquence 5 présentant un homme au téléphone marchant dans son jardin.

Méthode de spatialisation

L'expérience portant sur la cohérence de positionnement d'objets sonores à l'image, s'est posée la question de la méthode à employer pour spatialiser les sons. Protools HD proposant des bus au format 7.1 SDDS, il est tout à fait envisageable d'effectuer tous les mouvements et placements manuellement durant le mixage. De plus, différents outils de plus en plus performants sont à disposition des mixeurs. Il existe par exemple un plugin "Spanner" qui donne la possibilité de voir en surimpression sur la vidéo de protools, l'endroit où nous sommes en train de spatialiser la source. Cela permettrait donc de spatialiser assez précisément des sources sonores. Cependant cette méthode laisse trop de place à l'approximation humaine du mixeur.

Pour le cadre de cette expérience lui a donc été préféré un patch Max Msp permettant de spatialiser un son monophonique sur cinq canaux (grâce à une loi tangentielle classique) en renseignant le positionnement des enceintes, cette spatialisation se fait en pointant avec la souris la source sonore dans l'image. Si la démarche semble assez compliquée puisqu'impliquant de faire différents exports successifs et utilisant un outil peu commun, cela permet d'avoir une spatialisation précise et prenant en compte le placement des enceintes [7].

Les séquences ont donc d'abord été mixées en 5.1 classique puis un export a été fait de l'objet sonore à spatialiser dans le patch max/msp. Le problème de cette méthode est de ne pas pouvoir écouter directement le résultat de la latéralisation des sons. C'est pourquoi lors du mixage, régulièrement je faisais des essais de spatialisation afin de me rendre compte de l'effet rendu, bien que celui-ci soit à cette étape imparfait.

Gestion des Réverbérations

De nos jours, il est très fréquent d'utiliser des réverbérations artificielles sur les directs. Très souvent ces réverbs sont monophoniques s'il s'agit de raccorder un double ou un hf avec l'acoustique naturel du direct, mais elle peuvent également être en multicanal si l'idée est d'augmenter la sensation d'espace du lieu. Cependant, si la source bouge à l'écran et que le son le suit, comment doit-on gérer la réverb ? Plusieurs solutions sont envisageables. On peut faire suivre la réverb mono également, mais cela pourrait enlever la sensation d'espace amenée par la réverbération. On peut également envisager une réverb stéréo (ou multicanal) pour garder la sensation d'espace. De plus, on peut utiliser une fonction de "follow pan" sur l'envoi auxiliaire de la réverb ce qui permettrait de mieux coller avec les déplacements de l'objet sonore. C'est la solution qui semble la plus adaptée à un mixage classique.

Cependant, dans le cas de cette expérience, la latéralisation se fait à posteriori, et le patch Max/Msp fournit un fichier avec cinq pistes correspondant aux cinq canaux frontaux. La

7. Bien évidemment dans le cadre d'un mixage normal, les outils classiques peuvent déjà faire un travail précis, le contrôle du mixeur permettant d'obtenir des résultats satisfaisants.

tâche est donc plus compliquée. J'ai donc fait le choix d'utiliser une réverbération stéréo, et de ne pas utiliser la fonction follow pan. Ainsi, la réverbération bien que ne suivant pas les mouvements, apportera néanmoins un liant au mixage, et de plus ce sera la même réverbération que pour la version mixée au centre.

ANNEXE 5 : Interface graphique pour les sujets du test

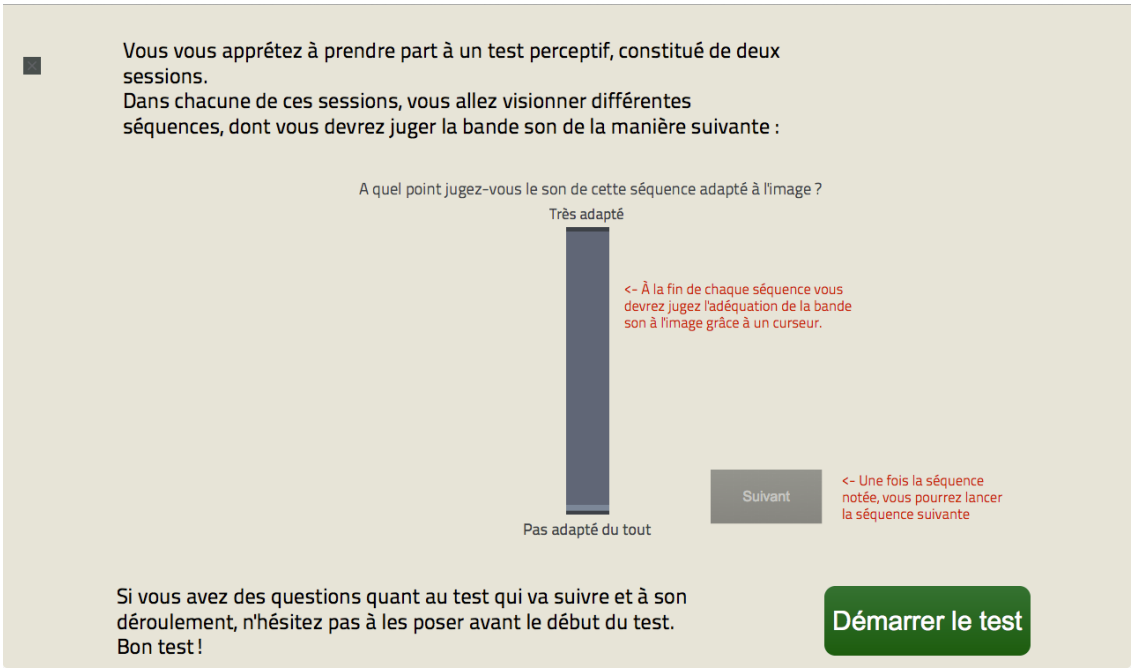


FIGURE A.7 – *Page d'accueil du test perceptif*

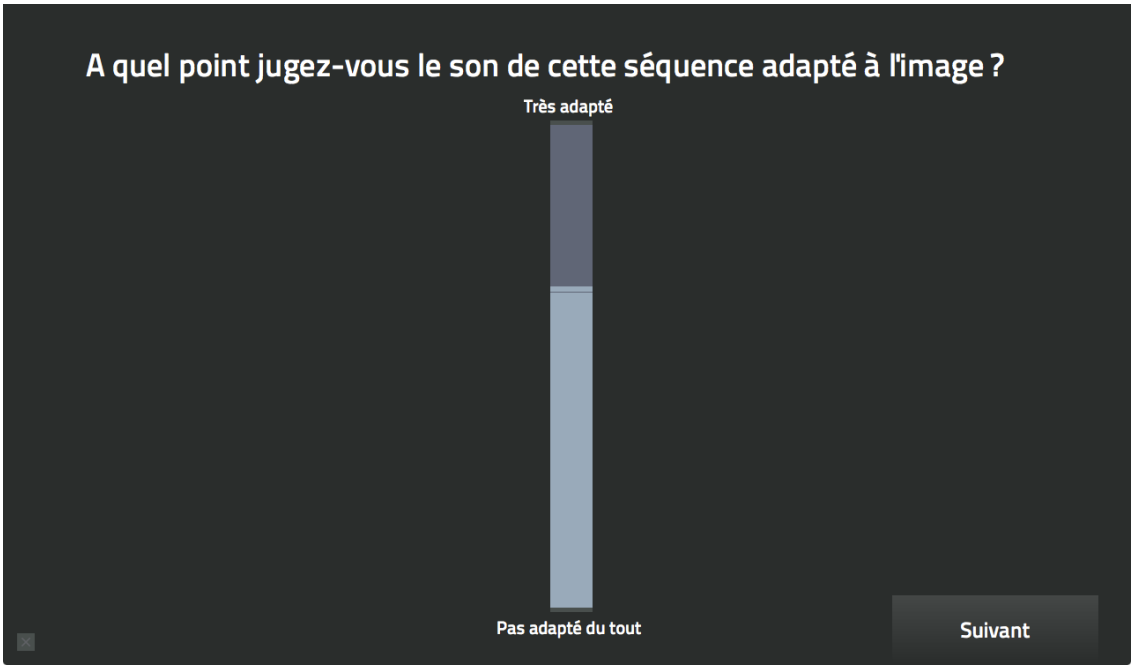


FIGURE A.8 – *Interface de notation du test perceptif*

FIGURE A.9 – *Page signifiant la fin du test perceptif*

ANNEXE 6 : Questionnaire post-test

Questionnaire post-test

Et non ce n'est pas encore terminé ! J'ai besoin de quelques informations concernant ta personne, mais aussi la manière avec laquelle tu as vécu l'expérience.

Informations personnelles :

- Nom : .....

- Prénom : .....

- Age : .....

-Profession/ Etudes/ Occupation : .....

- Fréquentation de salle de cinéma (entourer la bonne réponse):

0→2 par an / 3→6 par an

6→12 par an / plus de 12 par an

+ de 1 par semaine

Concernant l'expérience :

- As-tu ressenti de la fatigue durant le test ?.....

- Durant le test, as-tu perçu des mixages différents pour une même séquence ? OUI / NON

- Si oui, quelles étaient selon toi les différences entre les deux mixages ?

….....

.....

.....

.....

.....

- As-tu ressentie une gêne concernant le mixage de certaines séquences durant le test ? Et si oui peux tu expliquer ce qui t'as déranger ?

….....

.....

.....

.....

.....

FIGURE A.10 – *Première page du questionnaire post-test.*

- As tu des remarques concernant le test de manière générale, les séquences présentées, ou la bande son de ces séquences ?

….....
.....
.....
.....
.....

Merci pour ta participation !!

FIGURE A.11 – *Deuxième page du questionnaire post-test.*

Bibliographie

Articles et publications :

[1] ALAIS D. ET BURR D. (2004). "The ventriloquist effect results from near-optimal bimodal integration." *Current Biology*, 14 :257–262.

[2] ALLEN I. (1991). "Matching the sound to the picture.". *In Proceedings of the 9th Audio Engineering Society International Conference : Television Sound Today and Tomorrow.*

[3] 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*

[4] BATTAGLIA P., JACOBS R., ET ASLIN R. (2003). "Bayesian integration of visual and auditory signals for spatial localization". *Journal of Optical Society of America*, 20 :1391–1396.

[5] BERKHOUT A. J. (1988). "A holographic approach to acoustic control." *Journal de l'Audio Engineering Society*, 36 :977–995.

[6] BERTELSON P., RADEAU M. (1981). "Cross-modal bias and perceptual fusion with auditory-visual spatial discordance." *Perception & Psychophysic*, 29 :578–584.

[7] CHOE C. S., WELCH R. B., GILFORD R. M. ET JUOLA J. F. (1975). "The "ventriloquist effect" : Visual dominance or response bias ?" *Perception & Psychophysics*, 18 :55–60.

[8] COMOLLI J.-L., (2015). "L'oral et l'oracle, séparation du corps et de la voix" *IMAGES documentaires* "La voix", 55/56 :13-38.

[9] DOLBY LABORATORIES (2012). *Dolby Atmos : un son de prochaine génération pour le cinéma.* Document de présentation technique du système Dolby Atmos.

[10] FRISSEN I., VROOMEN J., DE GELDER B. ET BERTELSON P. (2003). "The afereffects of ventriloquism : Are they sound-frequency specific ?" *Acta Psychologica*, 113 :315-327.

[11] 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 Engineering Society Convention*

[12] JACK C. E. ET THURLOW W. R. (1973). "Effects of degree of visual association and angle of displacement on the "ventriloquism" effect." *Perception and Motor Skills*, 37 :967–979.

[13] KOMIYAMA S. (1989). "Subjective evaluation of angular displacement between picture and sound directions for HDTV sound systems." *Journal of Audio Engineering Society*, 37 :210–214.

[14] 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 Audio Engineering Society Convention*

[15] LEWALD J., EHRENSTEIN W. H. ET GUSKI R. (2001). "Spatio-temporal constraints for auditory-visual integration." *Behavioural Brain Research*, 121 :69–79

[16] PICK H. L., WARREN D. H. ET HAY J. C. (1969). "Sensory conflict in judgments of spatial direction." *Perception & Psychophysic*, 6 :203–205.

[17] RADEAU M. ET BERTELSON P. (1976). "The effect of a textured visual field on modality dominance in a ventriloquism situation." *Perception & Psychophysics*, 20 :227–235.

[18] RADEAU M. ET BERTELSON P. (1977). "Adaptation to auditory-visual discordance and ventriloquism in semirealistic situations." *Perception & Psychophysics*, 22 :137–146.

[19] SIPIÈRE DOMINIQUE (2011). "La voix au cinéma : divorces et retrouvailles" *Tropisme* "Questions de voix", 17 :159-172.

[20] THURLOW W. R.,ET JACK C. E.(1973). "Certain determinants of the « ventriloquism effect »." *Perceptual and Motor Skills*, 36 :1171–1184.

[21] VROOMEN J., BERTELSON P., DE GELDER B. (2001). "The ventriloquist effect does not depend on the direction of automatic visual attention." *Perception & Psychophysics*, 63 :651-659.

[22] WALLACE M., ROBERSON G., HAIRSTON W., STEIN B., VAUGHAN J. ET SCHIRILLO J. (2004). "Unifying multisensory signals across time and spac." *Experimental Brain Research*, 158 :252–258.

[23] 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." *Perception & Psychophysics*, 30 :557–564.

[24] WEERTS T. C. ET THURLOW W. R. (1971). "The effect of eye position and expectation on sound localization." *Perception & Psychophysics*, 9 :35–39.

Livres ou contribution à un ouvrage collectif :

[25] BAIBLÉ C. (1998). "L'image frontale, le son spatial". In *Cinéma et dernières technologies* dirigé par Frank Beau, Philippe Dubois et Gerard Leblanc, INA et De Boeck et Larcier, p.225-249.

[26] BOILLAT A. (1998). *Du Bonimenteur à la voix-over*, Antipodes, 2007.

[27] CHION M. (2005). *La voix au cinéma.*, Éditions de l'étoile/ Cahiers du cinéma (1982), réédition de 2005.

[28] COUTANT P.-A. (1991). *La reproduction du son au cinéma*. Femis, CST.

[29] KERINS M. (2011). *Beyond Dolby (stereo) Cinema in the digital age.*,Indiana University Press.

[30] VROOMEN J., DE GELDER B. (2004). "Perceptual Effects of Cross-modal Stimulation : Ventriloquism and the Freezing Phenomenon". In *Handbook of multisensory processes* dirigé par G. Calvert, C. Spence, et B. E. Stein, MIT Press, p.141-150.

Thèses et mémoires :

[31] CARREAU R. ET MACQUART T. (2015). *Utilisation de la technologie WFS dans la création sonore cinématographique : possibilités et limites*. Mémoire de fin d'étude de l'École Nationale Supérieure Louis Lumière.

[32] HENDRICKX E. (2015). *Cohérence des systèmes de diffusion sonore appliquée au cinéma en 2D et en 3D*. Thèse de doctorat, Université de Brest.

[33] MOULIN S. (2015). "Quel son spatialisé pour la vidéo 3D ? Influence d'un rendu Wave Field". In *Handbook of multisensory processes*, Thèse de doctorat, Université Paris Descartes.

Sitographie :

[34] DONNELLY C.(2014). *Dialogue on the move – panning in Gravity, Cars and Strange Days* article publié sur Designingsound.org
www.designingsound.org/2014/01/dialogue-on-the-move-panning-in-gravity-cars-and-strange-days/

[35] AMET H.(1911). *Method of and means for localizing sound reproduction*, brevet technique, disponible sur Google Brevets.
http ://www.google.com/patents/US1124580