Tifanie Bouchara

Stage Mars-Juil. 2008

M2 ATIAM (2007-2008)

Master de Sciences et Technologies de l'UPMC ;spécialité Mécanique et Ingénierie des Systèmes
Parcours Acoustique Musicale :Acoustique, Traitement du Signal et Informatique Appliqués à la Musique

RENDUS ANIMES NON AUDIO-REALISTES

Encadrement :

Christian JACQUEMIN, LIMSI-CNRS
Brian KATZ, LIMSI-CNRS

Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur, BP 133, 91403 ORSAY CEDEX
christian.jacquemin@limsi.fr ; brian.katz@limsi.fr

« We don’t see things as they are, we see things as we are »
Anais Nin

Remerciements

Je tiens à remercier tout particulièrement mes tuteurs de stage, Christian Jacquemin et Brian Katz, qui ont accepté de me reprendre sous leurs ailes et m'ont fait confiance une fois de plus. Merci donc pour m'avoir proposé un sujet aussi riche, pour avoir su m'encadrer de bons conseils tout au long de ce stage et pour m'avoir mené jusqu'à ma première conférence.

Merci aussi à P. Le Quéré, à Christophe d'Alessandro et J-P. Sansonnet pour m'avoir accueillie au LIMSI et dans leur équipe de recherche.

Un grand merci également à Mathieu Courgeon pour ses talents de « débuggeur » qui m'ont été d'un grand secours, et pour son aide en informatique graphique dont il a su partager les connaissances et un peu de sa passion. Un merci tout particulier à Marc Rébillat pour ses paroles réconfortantes, ses remarques toujours constructives et son aide en matière d'acoustique. Je remercie également Rami Ajaj pour son soutien constant et ses bonnes idées en algorithmie.

Merci aussi à tous les stagiaires de la salle 209 pour la bonne entente et la complicité installée qui ont rendu ce stage d'autant plus agréable.

Enfin, merci à tous ceux qui m'ont apporté une aide souvent précieuse et que je n'ai pas cité de peur de ne remplir les prochaines pages que de noms...

Résumé

En synthèse d'image numérique, le Non Photoréalisme s'appuie sur une grande variété de rendus et d'animations qui ne cherchent pas à reproduire ni l'aspect, ni le comportement des objets du monde réel. Le but du stage est d'étendre ces techniques au domaine du son en proposant des rendus Non Audioréalistes, dans un cadre audiovisuel. Nous nous sommes concentrée sur les méthodes de distorsions de l'espace focus+contexte utilisées en visualisation de données. Nous avons ainsi implémenté deux méthodes d'exploration d'espaces multimédias. En particulier, à partir d'un dispositif pour la création d'environnements multimodaux interactifs spatialisés, nous avons développé une lentille grossissante qui agit sur les deux modalités auditives et visuelles de façon simultanée et cohérente. Pour la modalité visuelle, elle s'apparente aux distorsions de type FishEye. Afin de tester cette nouvelle méthode d'interaction, nous avons mis en place un protocole expérimental basé sur des méthodes d'évaluations d'interfaces zoomables.

Mots-clés : scènes multimédias, rendu audio temps réel, audio 3D, rendu graphique interactifs 3D, techniques focus+contexte

Abstract

In computer graphics, non photorealistic rendering focuses on a wide variety of graphical animations and renderings that do not try to reproduce the look and the behavior of real world objects. Our aim is to extend these techniques to non audiorealistic rendering for sound in audiovisual context. We focused on space deformation with focus-and-context techniques used in the research field of visualization. We worked on two navigation methods for 3D multimedia environments. We also used a framework for the design of such multimodal interactive spatial virtual spaces. A new magnifier lens was developed to apply distorsion on both visual and audio modalities at the same time. This lens is similar to a FishEye view for its visual component. A user evaluation study based on multiscale evaluations methods was carried out.

Keywords : Multimedia scenes, Real-time audio, 3D audio, Interactive 3D graphics, Focus-and-context techniques.

Présentation de l'organisme d'accueil

Mon stage a eu lieu au **L**aboratoire d'**I**nformatique pour la **M**écanique et les **S**ciences de l'**I**ngénieur. Le LIMSI, unité propre du CNRS (UPR 3251), est situé à Orsay en région parisienne (Essonne, 91). Associé aux universités Paris 11 (Paris-Sud) et Paris 6 (Pierre et Marie Curie), il compte environ 120 permanents (chercheurs, ingénieurs, techniciens et administratifs) et une soixantaine de doctorants.

Le laboratoire est divisé en deux départements qui incluent 7 équipes, chacune s'occupant d'un sujet précis :

Département Communication Homme -Machine (CHM)

AMI --- Architectures et Modèles pour l'Interaction
LIR --- Langues, Information et Représentations
TLP--- Traitement du Langage Parlé
PS --- Perception et Située

Département Mécanique et Energétique (MECA)

AERO --- Aérodynamique instationnaire
CORO --- Convection et Rotation
TSF --- Transferts Solide -Fluide

Ainsi les thèmes de recherche menés au LIMSI couvrent un large spectre disciplinaire, allant de la « thermodynamique au cognitif », en passant par la mécanique des fluides, l'énergétique, l'acoustique, l'analyse et la synthèse vocale, le traitement de la langue parlée et du texte, la vision et la perception, la réalité virtuelle et augmentée.

J'ai été co-encadrée par deux personnes : Christian Jacquemin qui fait partie du groupe AMI (Architectures et Modèles pour l'Interaction) et Brian Katz qui travaille dans le groupe PS (Perception Située). Mon stage a donc pris place au sein de ces deux équipes de recherche.

Le groupe Architectures et Modèles pour l'Interaction (AMI)

Le groupe de recherche Architectures et Modèles pour l'Interaction (AMI) du département CHM du LIMSI-CNRS a été créé en 2001. Ce groupe a pour objet d'étude l'interaction pour elle-même, dans les systèmes d'information médiatisés par ordinateur. Par la provenance de ses membres, ainsi que dans les thèmes abordés, le groupe AMI affiche une volonté de pluridisciplinarité autour d'un même objet. A l'intérieur de ce groupe, j'ai travaillé sur le thème « Virtualité, Interactivité, Design & Art » (VIDA). C'est un thème de l'action transversale VENISE (Virtualité et Environnement Immersif pour la Simulation et l'Expérimentation). Il a pour but de promouvoir les recherches associant scientifiques, industriels, artistes et designers sur des projets en réalité virtuelle ou augmentée, graphique ou audio. Il est placé sous la responsabilité de Christian Jacquemin.

Le groupe Perception Située (PS)

Le groupe Perception Située a pour objectif l'étude de la perception dans ses diverses modalités (visuelle, auditive et gestuelle) pour le développement de systèmes de perception artificielle et d'interfaces expressives. Il y a trois axes majeurs dans ce groupe:

- Vision et robotique autonome (reconstruction d'images radiologiques 3D, représentation d'itinéraires et orientation d'agents autonomes) ;
- Traitement du langage (analyse, perception et synthèse de voix expressive);
- Réalité virtuelle (analyse et perception de scènes visuelles, perception et synthèse du son dans l'espace).

C'est sur cette dernière catégorie que mon stage s'est porté : le thème «Son et Espace». Ce thème est dirigé par Brian Katz. Il traite des aspects spatiaux de l'audition pour la réalité virtuelle audio (appelée aussi audio 3D), de l'acoustique dans l'espace (acoustique des salles, interaction salle/instrument, rayonnement des sources sonores) et d'acoustique des instruments de musique.

Tables des matières

Introduction

Depuis plusieurs années les recherches menées en informatique graphique ne se limitent plus au développement de rendus fidèles à la réalité, mais cherchent au contraire à développer d'autres modes de rendus visuels dits Non Photoréalistes (**NPR**). Ces rendus peuvent être de types très variés (imitations de styles, perspectives non-linéaires, rendus animés ou non…) et trouvent plusieurs applications qu'elles soient artistiques ou techniques. Parmi ces techniques, nous nous sommes focalisés sur celles exploitées dans le champ scientifique de l'exploration de données. Ces techniques combinent, à la base, visualisation et interaction, permettant un accès plus rapide à l'information. Or dans le cas de données multimédias, la composante sonore est un plus et les nouvelles approches d'exploration de données intègrent un rendu audio en complément de la visualisation graphique. Toutefois, aucune de ces méthodes de rendu audio ne se base sur un rendu non réaliste. Le but de ce stage est d'étendre les techniques NPR au domaine sonore. En particulier, nos recherches ont pour objectif d'optimiser des interfaces d'accès à des données multimédias, en proposant des outils combinant audio et graphisme. Parmi les différentes techniques déjà proposées dans le domaine de la visualisation, nous nous sommes appuyée sur les méthodes d'interfaces zoomables. Celles-ci sont basées sur des distorsions de l'espace de représentation et permettent d'associer des changements de perspectives ou même plusieurs perspectives différentes à la fois. Ce stage a ainsi eu pour but de reprendre certaines de ces méthodes, appelées méthodes « focus+contexte », et de les étendre à un contexte audio et audio-visuel, en proposant une interface multimédia.

Pour reprendre la répartition visuelle dans un espace 2D ou 3D des différentes informations, nous avons utilisé pour le rendu sonore des techniques de spatialisation du son. A partir de ces méthodes, nous proposons des transformations du rendu sonore afin de rester cohérent avec l'image.

Dans la première partie de ce rapport, nous étudierons les travaux réalisés précédemment dans le domaine du NPR, et nous présenterons des exemples de rendus NPR parmis les diverses possibilités que propose ce domaine. Afin de situer plus notre travail, nous présenterons également les méthodes issues du NPR qui sont utilisées en recherche d'information. Et pour étendre ces domaines à celui du son, nous présenterons également des techniques de spatialisation du son ainsi que les quelques propositions de distorsion de l'espace sonore que nous trouvons dans la littérature.

Dans un deuxième temps, nous présenterons le dispositif logiciel, constitué de deux parties, une pour le rendu sonore spatialisé et une pour le rendu graphique et la navigation, qui nous a permis de développer des environnements multimédias à explorer. Puis nous verrons comment à partir de ce dispositif, nous avons conçu un système de déformation de l'espace de type loupe grossissante, à la fois pour le rendu audio et le rendu graphique.

Enfin, afin d'évaluer le système conçu, une expérimentation a été mise en place. En s'appuyant sur des évaluations courantes dans le domaine des interfaces zoomables, cette expérience vise à comparer deux méthodes d'exploration d'espaces 3D qui ont été étudiées durant ce stage. L'une, plus basique, reprend une navigation par déplacement de caméra,

méthode dite Pan&Zoom, qui permet de garder un rendu audio-graphique cohérent et sans déformation, donc réaliste. L'autre méthode au contraire est une approche par lentille grossissante développée pendant ce stage, approche qui joue sur une distorsion de l'espace audio-graphique.

Nous présenterons dans une dernière partie les perspectives de recherches futures auxquelles nous avons pensées durant ce stage et qui pourront être, pour certaines, mises en place durant le dernier mois de stage.

I. Etat de l'art

Au début du projet, nous ne savions pas exactement quels traitements parmi les méthodes NPR pourraient être adaptées à l'audio. Un survol de ces techniques nous a permis d'avoir une idée plus précise des « effets » que l'on pourrait étendre aux rendus sonores afin que les traitements audio soient similaires ou du moins cohérents à ceux appliqués à l'image. Nous présentons donc ici les différentes techniques liées aux rendus non photoréalistes, à l'exploration de données ou encore au domaine de la spatialisation du son qui pourront permettre par la suite de développer une interface multimédia proposant des distorsions de l'espace.

I. 1 - Rendus Non Photoréalistes

I. 1. a - Définitions

Le terme **photoréaliste** est employé pour désigner des méthodes, techniques ou artistiques, par lesquelles on peut synthétiser des images suffisamment réalistes (Figure 1) pour être confondues avec des photographies. Au contraire, les techniques de Rendus Non Photoréalistes (**NPR** pour *Non Photorealistic Rendering*) ([12],[17]) ne s'appuient pas sur des phénomènes physiques pour imiter le « réel », mais sur des phénomènes perceptifs afin de remplacer, renforcer ou rediriger l'attention de l'utilisateur/spectateur (et bientôt auditeur), notamment en focalisant son attention sur l'information que l'on veut transmettre. C'est pourquoi les techniques NPR permettent de simplifier la scène, l'image ou d'accentuer certains éléments au détriment d'autres. Elles sont couramment utilisées dans les dessins techniques, en particulier en architecture ou en visualisation de données.

(a) **(b)**

Figure 1. (a) « The Classroom » et (b) « Main Street Blue » : Deux images de synthèses photoréalistes réalisées par Gilles Tran . Extrait de http://hof.povray.org/.

Il est nécessaire de comprendre comment fonctionne le système perceptif humain pour savoir quelles sont les informations que l'homme va naturellement percevoir, dans une image ou une scène sonore. A partir de ces données, on pourra appliquer des traitements déformants pour:

- renforcer des traits discriminants et caractéristiques et éliminer ce qui n'est pas essentiel pour la compréhension, comme dans un effet cartoon ou une caricature,
- ou au contraire, rendre mieux perceptibles des informations pertinentes, en affaiblissant des informations superflues, comme dans un effet de « zoom » sur certains éléments d'un environnement.

On peut finalement dire que le Non Photoréalisme se situe à la fois dans le domaine de la perception et dans celui de la représentation.

On peut distinguer deux grandes familles de méthodes NPR. La première famille regroupe un ensemble de techniques qui visent à transformer le contenu lui-même (changement de couleurs, de textures ou de forme) et qui en général cherchent à imiter des styles préexistants. Le deuxième groupe réunit des techniques qui agissent sur l'espace de représentation, en jouant en particulier sur la représentation de l'espace (perspective, lentille déformante, flou de profondeur...). Cette taxonomie en deux grands groupes se rapprochent de celle de Lansdown ([17]) qui distingue les méthodes qui peuvent être utilisées sur quasiment toutes les images de base, qu'il appellera alors « *Image Space Effect* » des méthodes qui nécessitent en général plusieurs images, il s'agit alors de « *Perspective Space Effect* ».

I. 1. b - Rendu NPR par imitations de styles

Les systèmes de rendus NPR présentés ici reprennent des styles déjà utilisés avant l'arrivée de l'informatique, d'où le terme parfois employé de **stylisation** ([8]). Il peut s'agir de techniques artistiques, liées au dessin (croquis) ou à la peinture (Figure 4). On peut alors obtenir une image générée par ordinateur à base de pinceaux, crayons, brosses ou plumes d'oies, exactement comme un artiste l'aurait fait à la main.

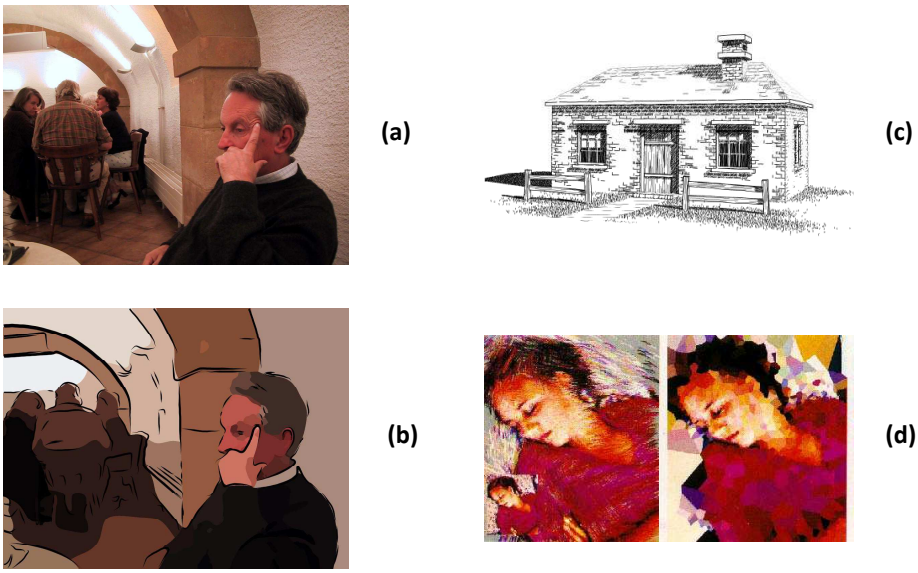


Figure 2. Exemple de rendus imitant des styles artistiques naturels générés automatiquement.
b) rendu stylisé à partir d'une photographie (a). c) Imitation d'un dessin à la plume par Winkenbach.
d) Rendus « impressionnistes » par P. Haeberli.

Pour parvenir à cela, on s'appuie sur les éléments essentiels d'une image. On jouera alors particulièrement sur les lignes de contours et les silhouettes. Ce genre de rendu s'appuie aussi sur des travaux du domaine de la perception des couleurs qui montrent que les couleurs chaudes et froides sont perçues très différemment. En plaçant deux sources de lumière, l'une chaude, l'autre froide (comme le soleil et le ciel), on a alors une meilleure

perception du volume 3D des objets, d'où l'utilisation de ces techniques en illustration pour réaliser des dessins techniques, dont un exemple est présenté en Figure 3 .

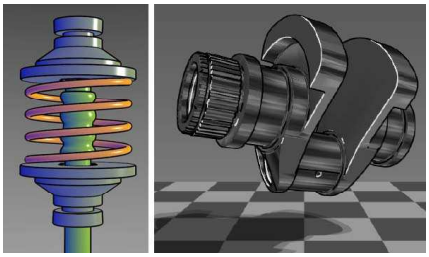


Figure 3. Techniques d'illustration non-photoréaliste avec modèle d'illumination non standard décrite par Gooch et al., 1998.

I. 1. c - Distorsion et perspectives

En adoptant d'autres perspectives que celle employée par le système visuel humain, telles que les perspectives parallèles, on parvient également à obtenir des rendus non photoréalistes très pertinents pour certains domaines, comme c'est le cas de l'architecture. Le cours de Kaleigh Smith ([25]) reprend en partie ces notions de jeux de perspectives.

Parmi ces rendus, on peut notamment trouver les projections non linéaires, comme celles que Coleman ([4]) utilise dans son film *RYAN* afin de présenter le point de vue subjectif de son personnage (Figure 3). On peut également jouer sur des perspectives multiples, qui au contraire des perspectives simples à une seule vue, permettent de capturer plusieurs perspectives en une seule image. Elles offrent donc une visualisation plus riche et complète où le champ visuel n'est plus restreint. Par exemple, en architecture ([24]), ces modèles sont utilisés pour donner une idée du rendu visuel que donneront tous les bâtiment d'une même rue. On obtient en effet une vue d'ensemble qui permet de voir les faces avant de tous les bâtiments simultanément(Figure 5).

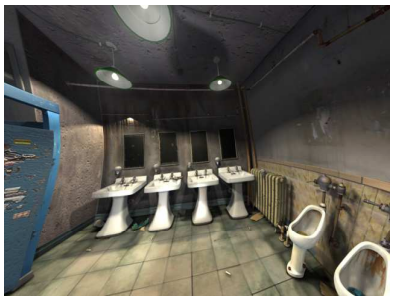


Figure 4. Vue subjective rendue par projection non linéaire. Extrait du film RYAN.

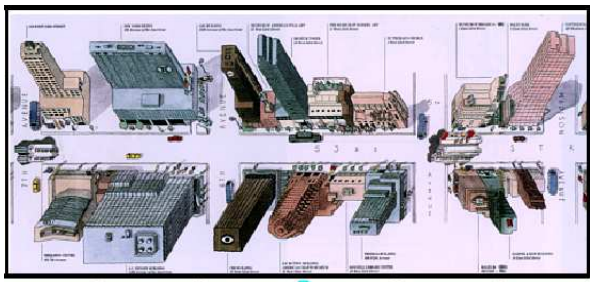


Figure 5. Perspectives à projection multiples créées à la main dans le cadre d'un projet architectural.

C'est sur ces méthodes par distorsion de l'espace que nous nous sommes focalisée durant ce stage. En effet, la distorsion de l'espace est très souvent appliquée au domaine de l'exploration de données car elle facilite l'accès à l'information.

I. 2 - Techniques de distorsion et d'interactions en visualisation d'informations

I. 2. a - Présentation générale

Dans le domaine de la visualisation de données, le but est de représenter un ensemble de données, en nombre important et liées par des relations complexes, dans un espace de représentation permettant d'explorer la base de données pour rechercher des informations précises. L'espace de rendu, essentiellement visuel, est limité (notion d'espace écran), il faut donc trouver des solutions pour afficher dans cet espace réduit un maximum d'informations, notamment un maximum de détails. De plus, l'utilisateur doit pouvoir se déplacer dans l'environnement de rendu afin d'accéder rapidement aux informations non visibles au départ. On parle d'**exploration de données** (*datamining* en anglais) ([16]). La visualisation/exploration des données doit se faire de façon interactive, c'est-à-dire réactive par rapport aux demandes de l'utilisateur ([15], [16]). Enfin, pour ne pas perdre celui-ci dans la masse de données à explorer, on utilise des techniques dites de distorsion ([20]). Elles présentent l'information sélectivement, soit en filtrant les données, soit en déformant l'espace de représentation.

Ces techniques de visualisation de données font appel à des interfaces homme-machine (IHM) très poussées qui sont absolument essentielles pour l'utilisation interactive en temps réel des outils proposés ([27]). Elles prennent ainsi en compte les contraintes humaines liées aux capacités motrices et cognitives des utilisateurs.

Un exemple de techniques de distorsion de l'espace se retrouve dans les **interfaces zoomables** ([19]), ou ZUI pour *Zoomable User Interfaces* (on parle aussi d'interfaces multi-échelles). Les données sont représentées grâce à une échelle de détail variable (zoom sémantique) en fonction de leur intérêt pour la tâche réalisée au moyen de cette interface. Il n'apparaît alors qu'une vue principale à la fois, appelée **focus**. D'autres techniques permettent aussi d'afficher une partie du **contexte** local (ce qui entoure spatialement le focus). On parle alors de **techniques focus+contexte** ([18]). Le contexte global (présentant la structure globale de l'espace de données) n'est pas représenté à moins de zoomer (vers l'arrière, on parle aussi de dézoom), auquel cas, c'est le focus qui n'est plus identifiable avec le niveau de détail souhaité. Les masses de données étant de plus en plus importantes, l'utilisateur se perd plus facilement lors de son exploration. Il est alors nécessaire de lui offrir une troisième couche (après le focus et le contexte), appelée couche historique. Elle permet à l'utilisateur de retourner aux régions déjà visitées dans l'espace d'information pour qu'il puisse relier la visualisation de ces régions avec le focus et la vue initiale. Nous ne nous intéresserons pour le moment qu'aux couches de focus et contexte.

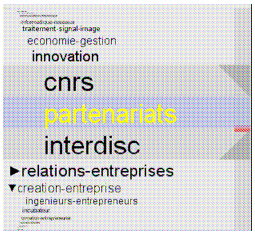


Figure 6. Exemple d'une liste zoomable par zoom sémantique avec une zone de focus et un contexte local.

I. 2. b - Technique Pan&Zoom

La méthode **Pan&Zoom** ([2],[13]) utilise une seule vue qui permet de voir tout le document avec une échelle uniforme. Le *Panning* permet de déplacer la vue sur les axes gauche-droite et haut-bas sans changer le niveau de détail. Au contraire, le *Zooming* permet de changer d'échelle : ainsi on peut tantôt s'éloigner des informations pour changer de centre d'intérêts tantôt s'en rapprocher pour se focaliser sur une partie du document.

Cette technique permet de se situer mais ne permet pas la visualisation simultanée du détail et du contexte. Pointer une cible par exemple avec cette technique nécessite 3 phases consécutives : d'abord on cherche une vue globale en dézoomant, puis on se déplace par panning pour se recentrer sur la cible et enfin on pointe en zoomant.

I. 2. c - Les lentilles magiques et l'affichage bifocal

Une autre technique utilisée en visualisation de données, ainsi que pour des rendus artistiques, est une méthode appelée « ***MagicLens*** » ([26],[28]). Cela consiste à déplacer une **lentille** (circulaire, rectangulaire ou cubique) sur le rendu visuel pour qu'à l'intérieur de cette lentille, le contenu change (Figure 7). On peut ainsi modifier la propriété des objets, notamment leur forme, leur couleur ou leur texture. Dans [28], les auteurs proposent également une lentille volumique qui permet par exemple de voir à travers des objets (**Figure 8**).

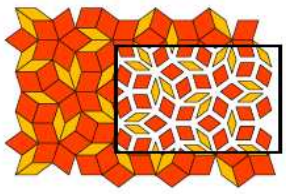


Figure 7. « MagicLens » 2D dite scalling lens. D'après un pavage de Doug Wyatt. Extrait de.

Figure 8. « MagicLens » 3D qui permet de donner un aperçu type rayon-X. Extrait de.

Parmi les types de lentille, nous nous intéresserons particulièrement aux lentilles grossissantes qui jouent sur un effet de profondeur et modifient l'échelle de détails à l'intérieur de la lentille. On parle d'affichage bifocal (« *Bifocal Display »)*. Ainsi, on peut visualiser deux vues à la fois : une vue avec un niveau de détail faible pour le contexte présentant ainsi une vue générale du document et une vue avec un niveau de détail plus fin sur une petite zone pour le focus. On parle aussi de techniques *Context+Detail*. Les deux vues sont superposées si bien que la vue « zoomée » masque une partie du contexte.

I. 2. d - Le système « Drag Mag »

La lentille *Drag Mag* est une alternative à la technique de *Bifocal Display*. Comme elle, la lentille *Drag Mag* permet de superposer deux vues d'un même document. La différence se présente dans le fait que la zone « agrandie » est déplacée et située dans le contexte : la zone est reliée au contexte par une vue en perspective (Figure 9). On peut alors comparer cela à un « gratte-ciel ». La base du bâtiment correspond à la zone initiale qui est recopiée avec un niveau de détail plus grand sur le toit (face avant). On peut alors déplacer la position de la lentille grossissante sur le document (base du « gratte-ciel ») pour explorer une autre

partie du document, mais aussi déplacer la projection de cette zone pour voir le contexte ou pour changer d'échelle de détail.

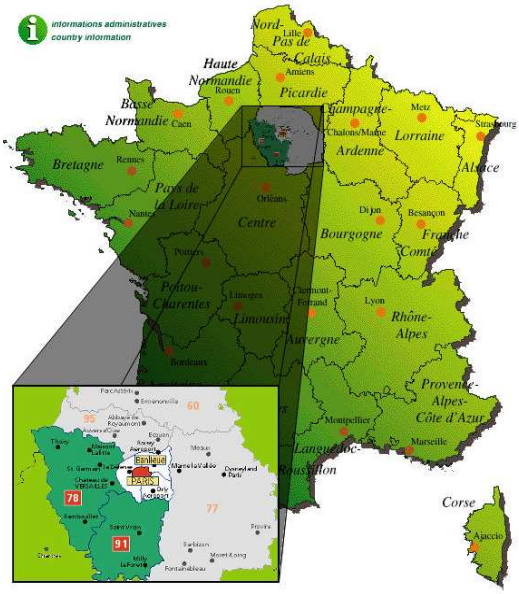


Figure 9. Vue en *DragMag* sur la région parisienne.

I. 2. e - Lentille en œil de poisson (« FishEye »)

Il s'agit d'une vue 2D déformée qui permet d'obtenir à la fois le contexte et le focus sur l'espace écran. Contrairement aux techniques « *Bifocal Display* », la transition se fait de façon continue entre le contexte et le focus.

Au premier plan (focus), les informations ont un degré de détails suffisant. Le reste de l'espace est en revanche d'une taille très inférieure et inversement proportionnelle à leur distance au centre de la lentille jusqu'à arriver à la zone extérieure de contexte (Figure 10).

(a) **(b)**

Figure 10. Effets de distorsion type loupe ou « *Fish-Eye* » par une déformation non linéaire (sphérique) (a, extrait de [4]) et par une déformation linéaire (b, extrait de [13])

Avec ce genre de méthodes, aucune information du contexte n'est masquée par la superposition de plusieurs vues. En revanche, la distorsion produite sur le bord des lentilles rend l'interprétation difficile pour cette partie du contexte.

Les techniques visuelles que nous venons de présenter s'appuient sur des déformations géométriques et des changements de point de vue ou d'angle de vue. La perception de l'espace est alors modifiée. Notre stage s'est focalisé sur l'extension de ces techniques uniquement, parmi toutes les techniques non photoréalistes, au domaine du son. Nous nous sommes donc intéressée aux méthodes de spatialisation sonore qui permettent de simuler un environnement sonore en trois dimensions en positionnant des sources sonores dans l'espace.

I. 3 - Quelques méthodes de spatialisation sonore

I. 3. a - La synthèse binaurale

Cette technique a pour but de reproduire le champ sonore tel qu'il serait perçu au niveau des oreilles de l'auditeur. Le rendu sonore est basé sur un traitement du son qui prend en compte les HRTFs (Head Related Transfert Functions). Cette notion de HRTF regroupe tous les modificateurs, différents pour chaque position de source, entre la source et nos tympans (diffractions sur notre torse et tête, masquage d'une oreille par la tête, filtrage fréquentiel des pavillons, etc…). Ce sont ces indices acoustiques que notre système auditif exploite pour la localisation sonore. En filtrant un signal monophonique par un couple de HRTF (oreilles gauche et droite), on génère une "source sonore virtuelle" que notre cerveau va placer dans la direction correspondante : c'est la synthèse binaurale.

L'écoute binaurale se fait sur casque. Idéalement, le traitement devrait prendre en compte directement les particularités morphologiques de l'auditeur. L'individualité des HRTFs rend la mise en œuvre de ce procédé assez complexe. De plus, bien que cette technologie restitue le son correctement dans tout l'espace, les ressources nécessaires pour un calcul en temps réel sont assez importantes et elles sont amplifiées dans le cas où un utilisateur est amené à se déplacer physiquement dans un environnement. Il faut alors envisager un suivi permanent de la tête de l'utilisateur. On parle de tracking.

I. 3. b - Le VBAP

Le *Vector Based Amplitude Panning* (VBAP) est une technique de spatialisation du son développée par Pulkki ([23]). Il s'agit d'une généralisation de la stéréophonie à un espace 3D. En stéréophonie, une paire d'enceintes permet de restituer des sources frontalement et dans le plan horizontal. Avec le VBAP, cette paire de haut-parleurs est remplacée par un triplet, ce qui permet de restituer des sources dans une partie plus large de l'espace et en particulier en élévation. En associant plusieurs de ces triplets, on parvient même à avoir une spatialisation en 3D tout autour de l'auditeur (périphonie). Comme en stéréophonie, la reproduction est limitée car elle n'est correcte que dans une zone restreinte appelée *sweet-spot*. Toutefois cette méthode, comme toutes les autres méthodes de stéréophonie généralisée, présente l'avantage de nécessiter peu de puissance de calculs et est facile à mettre en œuvre.

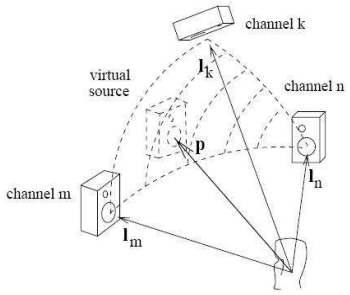


Figure 11. Principe de l'effet de sommation tridimensionnel, d'après [23].

I. 3. c - Holophonie : la Wave Field Synthesis

La synthèse de champ sonore (*Wave Field Synthesis* ou WFS) est une technique de restitution a priori physiquement exacte ([7]). Elle est basée sur le principe de Huygens-Fresnel : si l'on souhaite reproduire l'action acoustique d'une source sur un volume de l'espace, il suffit juste de savoir reproduire au niveau de la surface délimitant le volume étudié les ondes émise par la source. Donc le contrôle d'émetteurs acoustiques sur une surface délimitant un volume permet de reproduire un champ sonore donné au sein de ce volume.

Comme cette technique est plus complexe à mettre en œuvre que les précédentes, on limite en pratique, pour l'instant, la restitution des sources au seul plan horizontal. L'auditeur dispose alors d'une réelle perspective sonore au sein de laquelle il peut se déplacer librement car ce système de reproduction propose une zone d'écoute (ou "sweet spot") bien plus large que les autres techniques de reproduction et qui correspond au volume dans lequel le rendu est "exact". D'autre part la résolution spatiale des sources offertes par cette technique est très fine. Malheureusement, la WFS a un coût élevé tant d'un point de vue matériel qu'en termes de temps de calcul. Elle est donc très difficile à mettre en œuvre.

I. 3. d - Ambisonic

Le système **Ambisonic**, développé par Michael Gerzon ([10]), fut la première approche à chercher à reproduire le champ acoustique plutôt qu'à spatialiser de façon indépendante différentes sources sonores. Cela permit de sortir du plan horizontal pour chercher un rendu audio périphonique. L'approche ambisonique repose sur deux phases indépendantes. La première phase, dite d'encodage, est indépendante du système de restitution et repose uniquement sur la position des sources sonores. Elle est basée sur une décomposition en harmoniques sphériques. La seconde phase est une phase de décodage . Elle repose cette fois sur le nombre et la position des haut-parleurs pour la diffusion. L'utilisation de cette technique est relativement simple et permet de reproduire le champ sonore avec n'importe quel nombre de haut-parleurs (au moins 2) et n'importe quelle configuration, ce qui rend le système de diffusion modulable. Toutefois une répartition des haut-parleurs sur une sphère est conseillée (équidistance des HP par rapport à l'auditeur).

Un des gros avantages de l'Ambisonic est que c'est une technique qui ne demande pas plus de calculs quand le nombre de sources croît. En effet, la partie encodage, très faible en coût de calcul, ne demande pas de ressources supplémentaires quand le nombre de sources augmente. Quant à la partie décodage, elle ne dépend pas du nombre de sources mais du nombre de haut-parleurs . Ainsi l'Ambisonic est une technique qui convient très bien à un projet contenant beaucoup de sources sonores comme c'est le cas des environnements construits dans le cadre de l'exploration de données.

L'Ambisonic présente aussi des inconvénients, notamment au niveau de la précision. Pour l'augmenter, il faut notamment passer à des décompositions aux ordres supérieurs et placer plus d'enceintes. Or nous cherchons à implémenter un système où la précision à l'avant est bonne car on a veut dans un premier temps mettre en place une distorsion cohérente entre l'image et le son, donc dans le champ visuel à l'avant. Pour que la restitution soit correcte , il faut des haut-parleurs à l'avant ce qui signifie que les haut-parleurs masqueraient l'écran ou seraient masqués par l'écran.

I. 3. e - Technique mixte Ambisonic - binaural

La séparation entre l'encodage et le décodage dans la technique ambisonique permet de remplacer la phase de décodage ambisonique pour un rendu sur haut-parleurs par un décodage binaural pour une restitution au casque. On peut ainsi combiner les avantages de l'Ambisonic et du binaural, le temps de calcul reste faible même pour un très grand nombre de sources (encodage ambisonique) mais la restitution est plus précise qu'avec un rendu sur haut-parleur (décodage binaural).

Par ailleurs cette méthode combinant ambisonique et binaural est très pratique dans le cas où l'on veut « tracker » les rotations de la tête de l'utilisateur pour adapter l'espace sonore reconstruit. En effet, la représentation du champ sonore sous forme de canaux ambisoniques entre l'encodage et le décodage est puissante mais beaucoup plus simple (par exemple à l'ordre 1 on n'a que 4 canaux pour représenter l'ensemble des sources) et peut être modifié plus facilement. Ainsi, pour un rendu au casque avec tracking de l'utilisateur, en seulement quelques opérations sur les canaux ambisoniques, on peut faire tourner une scène sonore constituée d'une centaine ou plus de sources.

C'est pour ces raisons que nous avons adopté cette méthode mixte Ambisonic-binaural pour le rendu sonore spatialisé des environnements qui ont été développés durant ce stage.

I. 4 - Théorie de l'Ambisonic et distorsion de l'espace sonore

Il se trouve que très peu de techniques sonores cherchent à distordre l'espace. En effet toutes cherchent pour le moment à optimiser une reproduction fine et réaliste de l'espace. Toutefois une technique, appelée « *Dominance* » modifie, par un effet de renforcement de l'avant, l'espace sonore. Cette technique s'appuie sur l'Ambisonic.

I. 4. a - Précisions sur l'Ambisonic

L'encodage est ce qui permet de passer du système de prise de son à un ensemble de canaux ambisoniques représentant l'espace sonore de façon simplifiée. En fait cette phase d'encodage repose sur une décomposition en harmoniques sphériques du champ acoustique. Cette représentation est obtenue en écrivant l'équation d'ondes dans un système à coordonnées sphérique où un point de l'espace est décrit par sa distance au centre *r*, son angle azimutal θ et son angle d'élévation φ *(*Figure 12*).* Le problème de cette décomposition est qu'elle requiert un nombre infini d'harmoniques sphériques. Dans la pratique, on ne peut estimer, transmettre et exploiter qu'un certain nombre *M* de composantes. On tronque donc la décomposition à l'ordre *M*. Le champ de pression acoustique est alors exprimé, sous la forme d'une décomposition de Fourier-Bessel tronquée, par l'équation 1 qui implique $N = (M+1)^2$ composantes.

$$p(\vec{r}) = \sum_{m=0}^{M} i^m J_m(kr) \sum_{0 \le n \le m, \sigma = \pm 1} B_{mn}^{\sigma} Y_{mn}^{\sigma}(\theta, \varphi) \qquad \textbf{Équation 1}$$

Où : les Y_{mn}^{σ} sont les harmoniques sphériques (Annexe 1),

θ est l'azimuth (angle dans le plan horizontal),

φ le site (*elevation* an anglais, angle dans le plan vertical),

σ = ± 1, $i = \sqrt{-1}$,

mn est l'ordre de l'harmonique sphérique, $n \le m$,

Les $J_m(kr)$ sont les fonctions de Bessel sphériques de premières espèces (Annexe 1)

k est le nombre d'onde,

r est la distance entre l'origine du système de coordonnées et le point de mesure.

Chaque composante B_{mn}^{σ} est le canal ambisonique obtenu par projection orthogonale de la pression *p* sur l'harmonique sphérique correspondante.

Et M est l'ordre de la décomposition, de façon théorique, plus l'ordre *M* est grand et plus l'encodage est fin et précis.

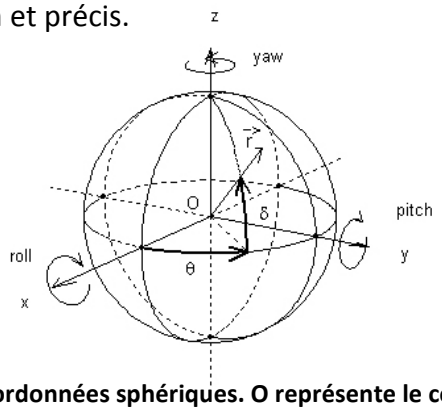


Figure 12. Système de coordonnées sphériques. O représente le centre du repère donc la tête de l'auditeur. On peut remarquer également les 3 degrés de rotation possibles. Extrait de [7].

La technique d'encodage consiste en fait à estimer la valeur des canaux B_{mn}^{σ} pour une source sonore *S* qui émettrait des ondes planes d'incidence (θ_S, φ_S), grâce à l'équation 2 ([6], p. 150) :

$$B_{mn}^{\sigma} = S.Y_{mn}^{\sigma}(\theta_S, \delta_S) \qquad \textbf{Équation 2}$$

Pour k sources sonores, on obtient ainsi, en superposant les canaux ambisoniques de toutes les sources sonores s_i, et en compactant l'ensemble des N canaux ambisoniques d'ordre M dans un vecteur :

$$\vec{B} = \sum_{i=1}^{k} \vec{Y}(\theta_i, \delta_i) \cdot s_i \quad ; \text{avec } \vec{B} = \begin{pmatrix} B_{00}^1 \\ B_{11}^1 \\ B_{11}^{-1} \\ \vdots \\ B_{M0}^1 \end{pmatrix} \text{ et } \vec{Y} = \begin{pmatrix} Y_{00}^1 \\ Y_{11}^1 \\ Y_{11}^{-1} \\ \vdots \\ Y_{M0}^1 \end{pmatrix}. \qquad \textbf{Équation 3}$$

La phase de décodage vise à reconstruire dans une zone d'écoute le champ acoustique précédemment encodé en fonction du système de diffusion ou d'écoute utilisé. Il s'agit donc de récupérer les signaux HOA (B_{mn}^{σ}), et d'obtenir à partir d'un matriçage de ces composantes un signal que chaque haut-parleur doit émettre (Annexe 2). Il est important de souligner que le nombre de haut-parleur nécessaires pour obtenir une bonne restitution (homogène sur tout l'espace) dépend de l'ordre *M* d'encodage. La technique permet une diffusion sur haut-parleurs avec des systèmes très divers. En pratique, il est fortement recommandé d'utiliser au moins *L=2M+2* haut-parleurs.

I. 4. b - Proposition de distorsion de l'espace sonore

Pour distordre la perspective sonore, une méthode, expliquée tout d'abord par Gerzon sous le terme « *dominance* » ([11]) a été reprise par Daniel sous le terme « distorsion de la perspective » ([6], p.166). Il s'agit d'une forme particulière de transformation de Lorentz qui permet, à l'ordre 1, de rendre les sons venant de l'avant (ou de la direction choisie) plus forts tandis que ceux de l'arrière sont diminués. Dans le même temps, la scène frontale est alors élargie ou resserrée. Cette méthode n'est présentée qu'à l'ordre 1 (on parle communément de B-format) et pour une transformation horizontale uniquement.

D'un point de vue mathématique, pour transformer le champ sonore encodé B en un autre champ ambisonique $\vec{B}'$, il suffit de multiplier $\vec{B}$ par une matrice de transformation T (matrice carrée de dimension NxN).

$$\vec{B}' = T\vec{B} \qquad \textbf{Équation 4}$$

Dans la déformation de perspective proposée par Daniel, le champ ambisonique $\vec{B}$ est, en fait, modifié en le multipliant par une matrice de transformation L_μ (L pour Lorentz). Ainsi, on obtient :

$$\vec{B}' = L_\mu \vec{B}, \quad \text{avec } B' = \begin{pmatrix} W' \\ X' \\ Y' \\ Z' \end{pmatrix}, \; B = \begin{pmatrix} W \\ X \\ Y \\ Z \end{pmatrix} \text{ et } L_\mu = \begin{pmatrix} 1 & \mu & 0 & 0 \\ \mu & 1 & 0 & 0 \\ 0 & 0 & \sqrt{1-\mu^2} & 0 \\ 0 & 0 & 0 & \sqrt{1-\mu^2} \end{pmatrix} \qquad \textbf{Équation 5}$$

La variable μ, appelée **coefficient de distorsion**, est comprise entre -1 et 1.

Ainsi, lorsque l'on applique cette matrice de transformations à l'ensemble des canaux ambisonique, on observe deux effets. Tout d'abord une **distorsion angulaire**, c'est-à-dire qu'une source initialement positionnée en (θ, φ=0) est déplacée en (θ', φ =0) tel que :

$$\theta' = \frac{\mu + \cos(\theta)}{1 + \mu \cos(\theta)}$$ **Équation 6**

Le déplacement se fait donc vers l'avant si μ>0, comme sur la figure Figure 13.

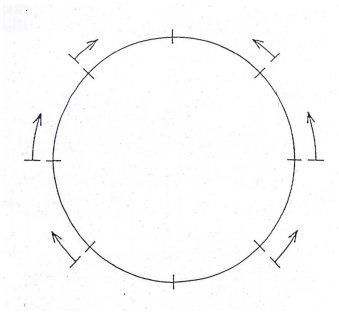


Figure 13. Schématisation de la distorsion angulaire dans le plan horizontal pour μ=1/3. Extrait de [11].

Dans le même temps, le gain de la source transformée est multiplié par un facteur γ :

$$\gamma = 1 + \mu \cos(\theta)$$ **Équation 7**

Si cette technique est la seule technique qui vise à distordre réellement l'espace sonore, elle est limitée cependant par deux aspects. D'abord elle modifie l'espace en entier et ne pourra pas être utilisée comme une lentille sonore où l'on appliquerait des transformations dans la zone à l'intérieure de la lentille tandis que ce qu'il y a autour (contexte) resterait intact. De plus cette méthode est limitée car on ne peut l'appliquer aux ordres supérieurs ([5]).

I. 5 - Conclusion partielle

Nous avons présenté quelques techniques de rendus visuels non photoréalistes. Parmi ces techniques, certaines permettent une distorsion de l'espace, en jouant sur les perspectives que ce soit en les assemblant ou en utilisant des non-linéarités. Elles sont alors très utiles en visualisation de données. Ce domaine cherche à rendre plus rapide et plus facile l'accès à une information contenue dans un ensemble de données très importante et s'appuie sur différentes méthodes de déformations et d'interactions. Dans le cas d'interfaces multimédias, la composante sonore, également importante, pourrait elle aussi être utilisée pour faciliter les recherches dans des espaces audiovisuels de données multimédias. Alors que plusieurs techniques ont été proposées pour la distorsion visuelle, seule une méthode a été proposée actuellement pour déformer l'espace sonore. Il s'agit d'une méthode de « Dominance » basée sur une technique de spatialisation du son Ambisonic. Afin de la mettre en place et de l'étudier dans un contexte audiovisuel, nous avons besoin d'outils pour créer des espaces multimédias. Nous pourrons alors proposer des méthodes, cohérentes dans le rendu audio-graphique, de distorsion de l'espace.

II. Conception d'un environnement d'accès à des données audiovisuelles

Nous avons vu dans la partie précédente diverses possibilités de rendus non photoréalistes qui visaient à distordre l'espace de représentations afin d'offrir de nouvelles possibilités pour l'exploration et la visualisation de données. D'autres part, nous avons vu différentes possibilités pour créer des espaces sonores en trois dimensions, espace que nous pourrons distordre comme le rendu graphique par la suite. Afin de combiner les deux modalités auditives et visuelles et d'obtenir des outils de navigation et d'interactions proposant une distorsion de l'espace, nous avons utilisé des outils audio et visuels qui permettent de créer des environnements virtuels en 3D.

II. 1 - Présentation générale

Lors d'un précédent stage au LIMSI, des outils pour la modélisation de scènes audiovisuelles en trois dimensions avaient été développés. Ces outils, regroupés dans un package appelé *SceneModeler* ([1]), avaient pour objectif de faciliter la création d'environnements interactifs temps réel, immersifs, à la fois sonores et visuels : des espaces navigables. Le package *SceneModeler* (son et graphique) est composé de deux parties : un descripteur de scènes virtuelles et un outil de spatialisation sonore. Deux logiciels sont utilisés : Virtual Choreographer (VirChor) pour la partie visuelle, Max/MSP pour le son. Tous deux gèrent le temps réel et permettent donc l'interaction que nous désirons mettre en place dans le cadre de notre recherche sur une optimisation des techniques d'exploration.

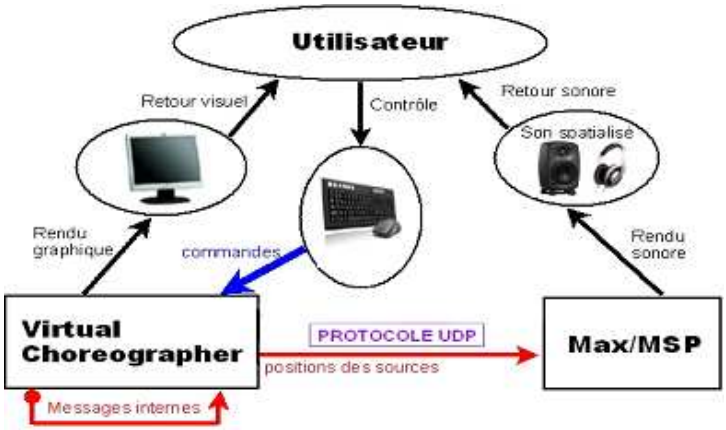


Figure 14. Architecture générale du dispositif utilisé.

L'architecture générale est présentée en Figure 14. D'un côté on a l'utilisateur, de l'autre la partie logicielle.

II. 2 - Modélisation 3D graphique interactive

La modélisation 3D se fait grâce à l'outil Virtual Choreographer (VirChor). Développé au LIMSI par Christian Jacquemin et ses collègues, il s'agit d'un navigateur de scènes 3D basé sur un langage type XML et s'appuie sur la librairie de rendu graphique multi-plateforme OpenGL. Il est distribué en Open Source sur http://virchor.sourceforge.net.

Le langage de description des scènes 3D dans VC est un langage XML qui décrit une scène par un graphe de scène (un treillis) composé de noeuds intermédiaires (des

transformations) et de noeuds terminaux (des objets géométriques ou non). La Figure 15 présente un exemple de ce type de graphe dont la partie en gras est codée en Figure 16.

Dans ce langage, on définit les propriétés géométriques et graphiques des objets comme la forme, la position/orientation, les couleurs. Chaque scène modélisée doit présenter au moins un utilisateur et une caméra (point de vue et point d'écoute par défaut de l'utilisateur). Le langage XML utilisé est très proche de celui de X3D, le langage du W3C pour la description de scènes géométriques 3D. Une différence majeure entre VirChor et les moteurs de rendu 3D temps réel, tels que Virtools ou des moteurs de jeux tels que Ogre3D, est que VirChor permet d'ajouter aux objets géométriques des propriétés sonores (type fichier de son émis, niveau, nombre de boucles, et bien sûr sa position liée au nœud). La scène 3D ainsi décrite possède donc plusieurs informations relatives à l'audio, qui peuvent ensuite être utilisées par des plugins ou transmises à d'autres applications par communication réseau , mais VirChor ne contient pas son propre moteur de rendu audio.

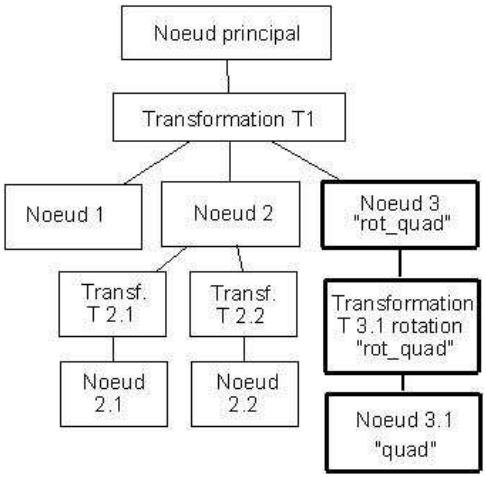


Figure 15. Exemple de graphe de scène

```
<node id = "rot_quad" >
<transformation id="rot_quad"
                geometry="rotation"
                angle = "50"
                x = "0" y = "1" z = "0"/>
        <node id = "quad">
             <use xlink:href = "quad.xml:#quad"/>
        </node>
</node>
```

Figure 16. Un exemple de code XML sous VirChor : le noeud rot_quad permet d'instancier un objet carré et de lui appliquer une rotation.

Comme la plupart des moteurs graphiques 3D, on peut définir les comportements des objets composant les scènes géométriques pour les animer en temps réel et présentent la possibilité de gérer des interactions soit entre l'utilisateur et la scène, soit entre les objets eux-mêmes. Cet aspect dynamique est intégré à la scène par des scripts (Figure 17). Cela fonctionne par envoi de messages entre les différentes composantes de la scène ou entre applications (typiquement par exemple entre VirChor et Max/MSP). Avec ce système de communication, le rendu visuel et sonore est calculé en temps réel via les deux logiciels.

```
<script id = "script key_capture">
    <command>
      <trigger type = "message_event" value = "key-f"
        state="active" bool_operator="=="/>
           <action>
                <set_node_attribute_value operator = "=">
                    <transformation angle = "({$root:angleyaw}+1)"/>
                </set_node_attribute_value>
              <target type = "single_node" value = "#camera rotation 1_Y"/>
             </action>
    </command>
</script>
```

Figure 17. Exemple de script et de commande : changement du paramètre "angleyaw".

II. 3 - Traitement de l'audio 3D

Dans notre système, pour le rendu sonore, nous avons utilisé Max/MSP. Il s'agit d'un environnement visuel pour la programmation d'applications interactives en temps réel. Max/MSP (Ircam/Cycling'74) est la combinaison du logiciel MAX pour le contrôle temps réel d'applications musicales et multimédias interactives par MIDI et de MSP, une bibliothèque d'objets pour l'analyse, la synthèse et le traitement du signal audio en temps réel. Chaque application Max/MSP, appelé patch, est écrite dans un langage graphique : on connecte entre elles des boîtes possédant des entrées et des sorties et jouant chacune un rôle précis. Max/MSP utilise la métaphore de modules électroniques que l'on branche entre eux par des câbles.

Dans le système utilisé, Max/MSP reçoit les données géométriques (azimut, élévation et distance des sources par rapport à l'auditeur – liés avec le point de vue « user » dans la modélisation géométrique) de chaque source, ce qui permet de leur appliquer un traitement sonore afin de simuler leur position dans l'espace.

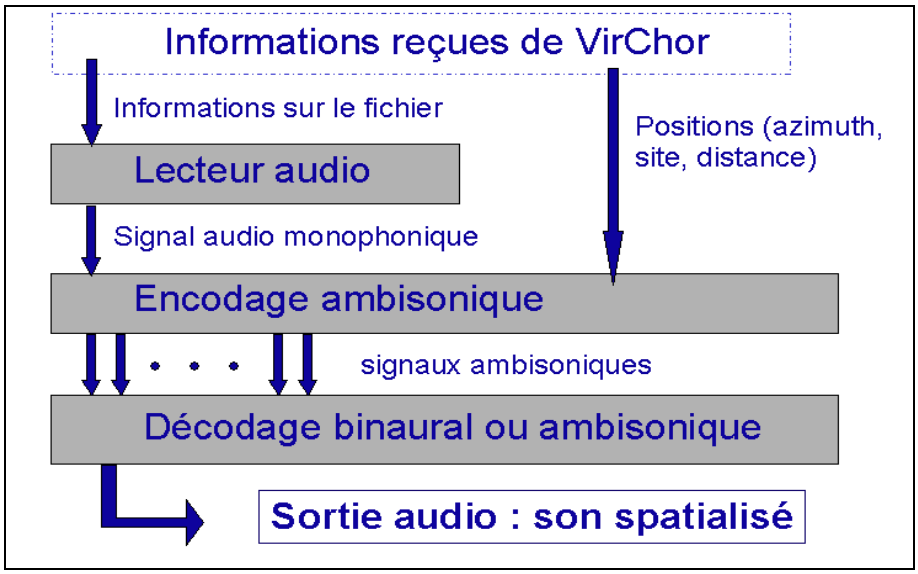


Figure 18. Schéma du processus de base de spatialisation des sons réalisés avec Max/MSP

Ainsi pour le rendu audio, nous fonctionnons en 3 phases (Figure 18, page précédente). D'abord, pour chaque objet visuel de la scène décrite dans VirChor, on associe un fichier son monophonique. La première phase consiste donc à déclencher la lecture de ce son qui sera ensuite encodé, par méthode ambisonique. Puis, en fonction du système de diffusion choisi (casque ou haut-parleurs), une phase de décodage permet de transformer les canaux ambisoniques obtenus en signaux sonores. On obtient alors une spatialisation des sons cohérente à l'image.

II. 4 - Communication inter-logicielle

Afin que VirChor puisse envoyer les informations nécessaires au calcul du rendu sonore, il est nécessaire de développer un système communicant entre les différentes applications. On peut choisir, par le système mis en place, soit d'utiliser un réseau local, en travaillant avec une seule machine, soit une architecture répartie sur deux ordinateurs. Lors de ce stage, nous avons décidé d'utiliser deux ordinateurs de façon à ce que l'un gère uniquement le rendu sonore, tandis que l'autre gère la navigation et le rendu graphique. Cela diminue ainsi les temps de calculs. Dans le dispositif utilisé ici, l'ensemble des communications réseaux entre les logiciels se fait par des messages textes au format OSC (OpenSoundControl) par protocole UDP (User Datagram Protocole).

II. 5 - Interaction homme-machine

Le système logiciel présenté peut être résumé par un schéma commun dans le domaine des interfaces homme-machine (IHM). Il s'agit du schéma modèle-vue-contrôleur (Figure 19). Ainsi d'un côté on a le modèle, avec la description de l'environnement, de l'autre la partie qui relie le modèle à l'utilisateur. Celle-ci est constituée d'une partie contrôle qui permet à l'utilisateur de gérer l'interaction (navigation) avec l'environnement et d'une partie rendu, ici visuel et sonore, qui donne un retour à l'utilisateur.

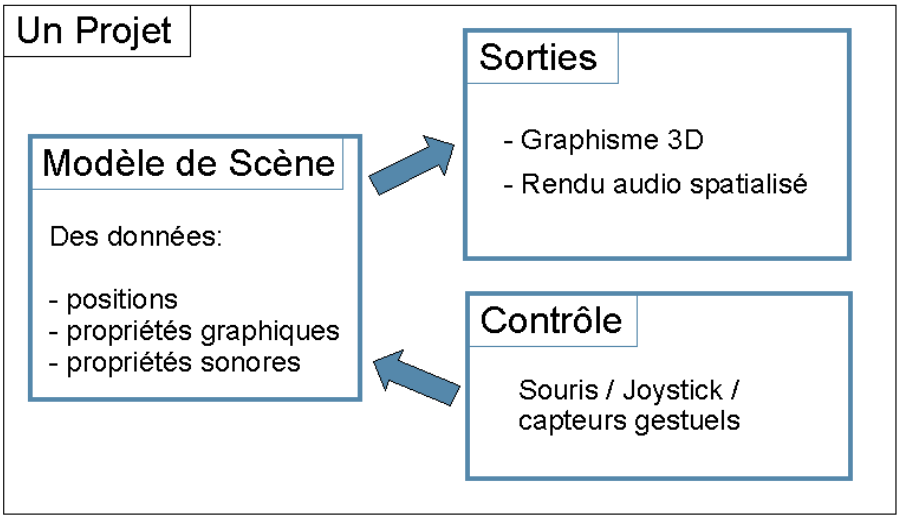


Figure 19. Présentation de principe de modèle-vue-contrôleur appliqué au dispositif mis en place pour la création et la navigation au sein d'environnements multimédias spatialisés.

III. Distorsion audio-graphiques dans une interface multimédia

Nous venons de présenter les outils, VirtualChoreographer et Max/MSP, par lesquels nous définissons des environnements multimédias spatialisés et interactifs. Nous présenterons dans cette partie les différentes techniques de distorsion que nous avons mis en place, à partir de ces logiciels. Nous commencerons par présenter les différents essais qui ont été faits durant ce stage avant de parvenir à une technique de distorsion de l'espace à la fois audio et graphique satisfaisante. Il s'agit d'une lentille FishEye étendue au domaine multimédia et non plus seulement visuel.

III. 1 - Distorsion visuelle

Les techniques de déformation du rendu visuel, qui ont été implémentées durant ce stage, ont été conçues en collaboration avec Mathieu Courgeon, stagiaire en informatique graphique au LIMSI.

III. 1. a - Etudes préliminaires : essais sur l'aspect graphique.

Affichage bifocal

Nous avons commencé dans un premier temps par mettre en place une lentille grossissante de type affichage bifocal (cf. I. 2. c - p. 14). Il s'agissait donc d'une déformation qui changeait l'échelle de détail pour une zone de l'écran prédéfinie correspondant au focus. Pour implémenter cette déformation linéaire, nous nous sommes appuyés sur un changement de perspectives à l'intérieur de la lentille. Il s'agissait en fait de donner pour cette zone de nouvelles informations sur la perspective virtuelle.

En effet, en informatique graphique, on définit la perspective, centrale ou ponctuelle, par ce qu'on appelle un frustum (un tronc de pyramide). Il s'agit d'un volume de vue associée à une projection en perspective dont le centre O est la position de l'observateur. Les objets contenus à l'intérieur de ce volume de vue sont projetés sur le plan de projection appelé plan *near* par projection linéaire tandis que les autres ne sont pas visibles. En modifiant les paramètres du plan *near*, on peut modifier la perspective du rendu donc la perception de l'espace visuel.

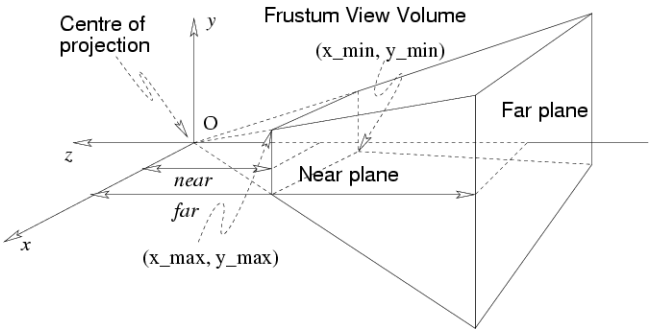


Figure 20. Volume de vue pour une projection centrale.

En particulier, si l'on diminue les dimensions du plan *near*, cela revient à augmenter la focale et les objets apparaissent plus grands comme vus au téléobjectif. C'est ce qui se passe exactement lorsque l'on zoom avec un objectif visuel : les objets paraissent plus gros mais surtout ils sont déformés, écartés vers l'extérieur, car l'angle de vue change.

Pour réaliser cette lentille type affichage bifocal, nous avons utilisé ce changement de perspective. La séparation entre les deux vues a été implémentée par un *shader*[1] utilisé pour changer la texture d'un rectangle placé exactement sur le plan *near* de façon à prendre tout l'espace écran. On appellera ce rectangle « rectangle de rendu » et la texture associée « texture de rendu ».

En fait, la scène est rendue deux fois. A chaque fois, le rendu graphique est stocké dans une texture, et ce pour chaque image du rendu. Le *shader* permet de sélectionner quelle partie de ces deux textures va être appliquée sur le rectangle de rendu et à quel endroit du rectangle : au centre de la lentille, on applique la texture correspondant au rendu zoomé de la scène, tandis qu'à l'extérieur, on utilise le rendu normal (Figure 21). Entre les deux zones, un flou blanc a été placé de façon à masquer le passage d'une texture à l'autre. De plus pour atténuer le passage brusque d'une échelle à l'autre, nous avons ajouté une zone d'interpolation très fine.

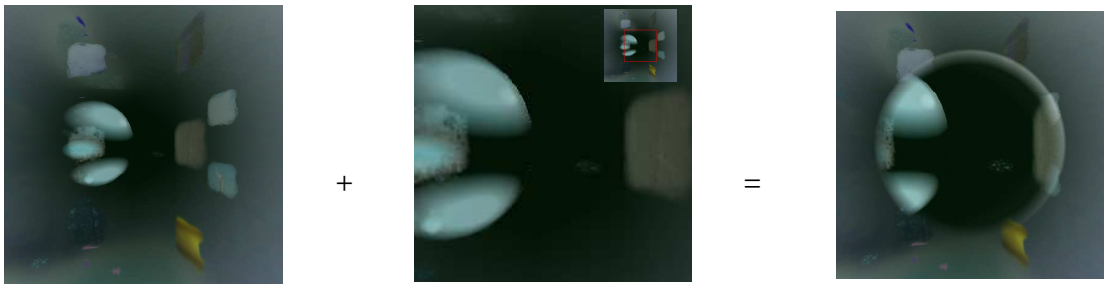


Figure 21. Rendu visuel (c) de la loupe implémentée par la méthode de *Bifocal Display*, présentant deux perspectives simultanées. Il est obtenu par superposition de deux textures de rendu, l'une normale (a), l'autre zoomée (b). La scène est celle créée pour le projet *MontheyO4* d'Antonio de SousaDias ([1]).

Cette technique présente plusieurs avantages. D'abord elle permet d'obtenir une vue zoomée non distordue et sans perte de qualité. De plus c'est un procédé que nous pourrons facilement réutiliser pour créer une lentille magique. En effet, au lieu de changer uniquement la perspective à l'intérieur de la lentille, nous pourrons tout à fait modifier également le contenu. L'inconvénient majeur de cette technique était lui prévu : les deux vues sont superposées donc le focus (intérieur de la lentille) masque une partie du contexte et l'on perd alors des informations.

Nous avons, pour éviter ce problème, cherché à développer une autre lentille visuelle, cette fois de type FishEye.

Optimisation

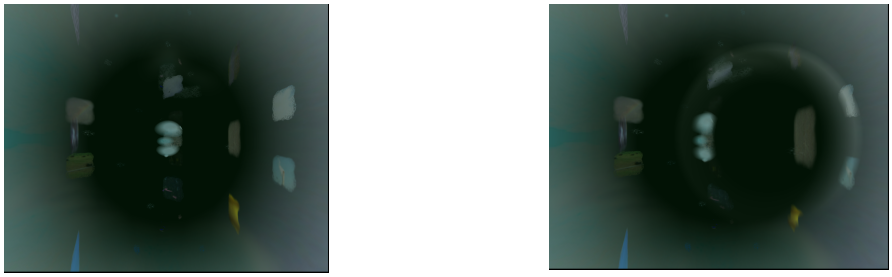
Notre deuxième version de lentille visuelle visait à combiner les avantages de la méthode précédente et des distorsions de type FishEye, sans pour autant en prendre les inconvénients.

Les vues *FishEye* habituelles se servent d'une déformation vectorielle qui a pour effet d'agrandir les pixels dans la zone de focus. Avec OpenGL, et par conséquence avec le logiciel

[1] programme permettant de réaliser une partie du processus de rendu directement par la carte graphique

de rendu graphique VirChor, on ne s'appuie cependant pas sur une définition vectorielle des images (comme en bitmap). Les déformations en *FishEye* appliquées directement sur la texture de rendu impliqueraient donc une perte de qualité car on verrait apparaître une pixellisation. Pour éviter, dans un premier temps, cet effet « pixel », nous avons repris la base de la lentille précédente qui permettait d'avoir un meilleur niveau de détail du focus, et ce sans aucune déformation ni perte. Nous l'avons modifié pour que les deux vues, focus et contexte, ne se masquent plus : nous avons alors utilisé un procédé pour passer d'une vue normale à une vue zoomée de façon progressive.

La technique que nous avons alors proposée effectue non plus deux mais six rendus successifs de la même scène avec des perspectives différentes. On peut alors parler de *Multifocal Display*. Les 6 rendus sont superposés par des cercles concentriques de plus en plus petits (Figure 22 b). Ainsi, sur les bords de la lentille les objets de la scène sont distordus de façon à compenser l'élargissement de la zone centrale de la lentille. On ne perd alors plus d'informations par masquage entre les différentes vues. L'échelle reste, de plus, homogène dans la zone centrale, ce qui facilite la compréhension.



(a) **(b)**

Figure 22. Captures d'écran d'une vue avant distorsion (a) : vue générale et de la superposition des 6 rendus pour la nouvelle lentille visuelle mise en place.

Cette méthode de rendu n'était toutefois pas satisfaisante et ce pour plusieurs aspects. D'abord, la jonction entre les différents rendus se faisaient mal, et ce malgré une interpolation entre les différentes zones. De plus, le calcul des 6 rendus successifs entraînait une charge de calcul trop forte pour la carte graphique ce qui était très coûteux en temps de calcul.

III. 1. b - La lentille grossissante visuelle finale

Les précédents essais de loupe visuelle étaient basés sur un modèle de déformation type *Bifocal Display*. Le système que nous avons finalement décidé de garder est lui basé sur un modèle en *FishEye.* La distorsion effectuée est une déformation sphérique.

Pour implémenter cette lentille, un autre *shader* a été programmé. Il modifie directement l'application de la texture de rendu sur le rectangle de rendu de sorte que les pixels sont étirés au centre de la lentille. En principe, du centre jusqu'au bord de la lentille, l'étirement des pixels se réduit pour revenir à une taille normale à l'extérieur de la lentille. En pratique, nous avons décidé de laisser une zone centrale avec une échelle constante. L'espace écran est alors délimité en trois zones :

- l'extérieur de la lentille est une vue large, avec un niveau de détail faible
- l'intérieur de la lentille est une vue finalement resserrée, avec un niveau de détail fin

- entre les deux, une zone permet de compenser l'effet de grossissement de l'intérieur de la lentille : les objets sont distordus.

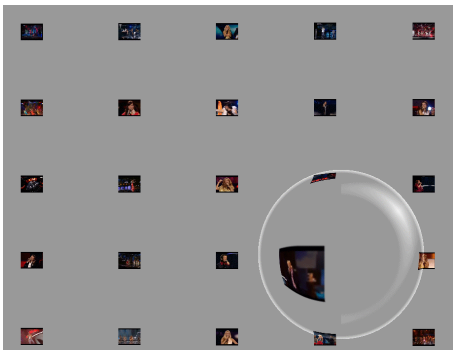


Figure 23. Rendu visuel obtenu avec le nouveau procédé de distorsion type *FishEye.*

Un reflet a été ajouté pour renforcer l'effet de relief de la lentille et un tour blanc permet de délimiter clairement le début de la distorsion et de pouvoir situer la lentille dans le cas de faible niveau de distorsion.

Cette méthode de déformation implique un effet non désiré de pixellisation mais le rendu est beaucoup plus clair et esthétique que dans la lentille précédente car il y a vraiment une continuité entre le focus et le contexte.

Pour modifier la déformation et accentuer ou diminuer l'échelle de détail, nous utilisons un coefficient de distorsion. C'est ce paramètre qui est envoyé au logiciel de rendu sonore, Max/MSP, pour obtenir une distorsion cohérente entre l'image et le son.

III. 2 - Distorsion audio

III. 2. a - Etudes préliminaires : les essais pour une lentille audio

Du côté de l'audio, nous avons testé, dans un premier temps, la seule méthode de distorsion de l'espace sonore que nous avions trouvée dans la littérature. Il s'agit de la méthode « Dominance » (cf. p. 20). Cette méthode a été très rapide à implémenter avec le dispositif de rendu sonore que nous avions mis en place. En effet, la distorsion proposée s'effectue sur les canaux ambisoniques obtenus après l'encodage. Ainsi, il nous a suffit de rajouter une étape de traitement entre l'encodage et le décodage (Figure 24). Le schéma du système de traitement du son 3D que nous avions développé auparavant (cf. Figure 18, p. 24) devient celui présenté en Figure 24.

Les résultats obtenus sont bien ceux qui étaient décrits par Gerzon et Daniel ([11], [6]), donc on a bien un écartement des sources vers les côtés, ainsi qu'un gain en niveau pour les sources placées à l'avant. Toutefois, cette méthode n'agit pas seulement sur l'avant mais sur tout l'espace sonore. Pour remédier à cela, nous avons pensé à compenser l'effet de gain sur l'arrière en appliquant manuellement un gain à la sortie sur les haut-parleurs.

Malgré tout, nous n'avons pu garder cette méthode qui est, d'une part limitée à l'ordre 1 ambisonique ce qui ne permet pas un rendu spatialisé optimal, et qui, d'autre part, ne permet pas de définir une zone unique de déformation comme c'est le cas avec une lentille.

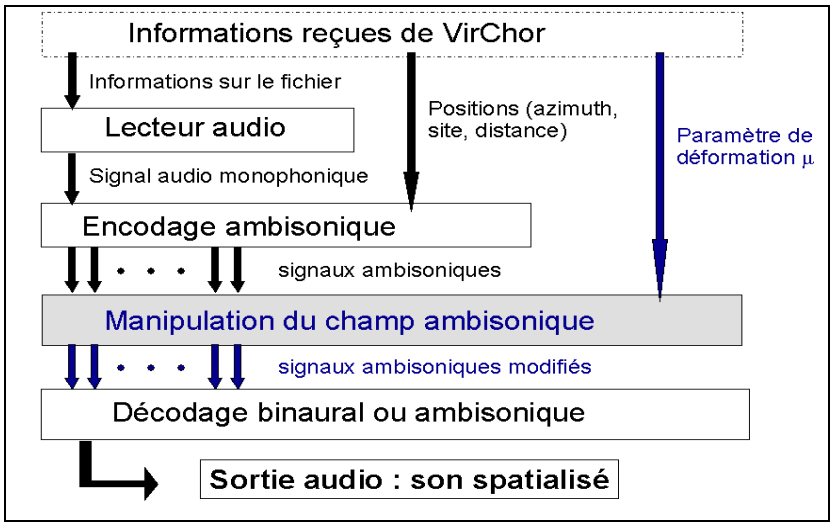


Figure 24. Schéma du processus du traitement des sons pour une transformation de type « *Dominance* » entre l'encodage et le décodage.

III. 2. b - Processus pour la distorsion audio

Nous cherchions à créer un système de distorsion audio qui serait cohérent avec la distorsion visuelle. De ce fait, nous avons calqué le processus de déformation de l'audio sur le rendu graphique. Dans un premier temps, nous avons dû déterminer si les objets étaient ou non situés dans la zone de focus. En effet, la distorsion visuelle est effectuée sur la texture de rendu directement et donc de façon indépendante des objets : même sur écran noir, la loupe apparaît. En revanche, le traitement sonore dépend lui des sources et de leur positionnement. En fonction des paramètres du volume de vue principale et de la lentille visuelle (position, dimension, coefficient de distorsion), nous regardons, pour chaque objet de la scène, s'il est projeté dans l'espace défini sur l'écran à l'intérieur de la lentille. Si c'est le cas, nous calculons alors, en temps réel toujours, la position que devrait avoir l'objet pour être projeté visuellement à cet endroit avec une vue normale sans distorsion. C'est cette nouvelle position, que l'on nommera « position apparente », qui est prise en compte pour rendre la spatialisation des sons. Nous obtenons alors un processus de la forme indiquée en Figure 25 :

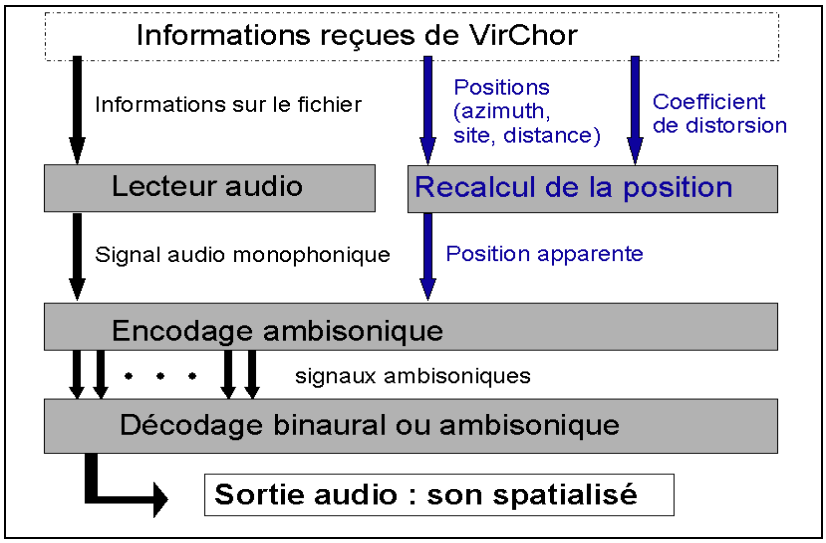


Figure 25. Processus de traitement sonore réalisé pour obtenir une lentille déformante sonore.

Cette méthode permet de rendre effectivement un rendu sonore cohérent avec le rendu visuel, ce que la méthode « *Dominance* » ne proposait pas. Cependant, alors que l'autre méthode agissait sur les canaux ambisoniques après l'encodage et ne dépendait donc pas du nombre de sources, la méthode que nous utilisons demande, pour chaque objet de la scène comprenant une composante sonore, des calculs supplémentaires. Plus la scène devient complexe et plus le temps de calcul augmente, ce qui n'est pas optimal.

III. 3 - Elaboration d'une interface d'accès à des données multimédias

Afin de concevoir, optimiser mais aussi expérimenter les outils et les modes de navigations proposés, nous avons pensé à des environnements multimédias particuliers. En effet, la lentille proposée devait être à la fois audio et visuelle. Pour bien nous rendre compte de la distorsion combinée, nous devions donc faire en sorte que la majorité des sources soient situées dans le champ visuel, d'où l'idée de proposer dans un premier temps des scènes où les objets seraient répartis de manière plutôt frontale.

De plus, nous voulions qu'il y ait suffisamment d'objets pour que les outils proposés aient de l'intérêt. En effet, s'il y a peu de données, tout tient dans l'espace écran à un niveau de détail suffisant pour la modalité visuelle et la distorsion n'apporte plus d'améliorations.

Enfin, la relation entre la partie visuelle et la partie sonore d'une source devait être bien marquée, condition nécessaire pour que l'on puisse percevoir la cohérence de position mais aussi de distorsion entre les deux modalités audio et visuelle. Or par expérience sur un ancien projet d'espace navigable appelé *Monthey04* ([1]), nous avions pu remarquer que la position spatiale ne suffit pas pour relier une source sonore à un objet visuel. C'est aussi ce qui ressort des études sur l'"effet ventriloque" : tant que les mouvements de la marionnette sont synchronisés avec les paroles du marionnettiste, alors le cerveau interprète ce qu'il entend comme provenant de la marionnette. Ces études ont aussi montré que le système auditif présente une certaine plasticité face au système visuel : il adapte la localisation perçue des sons en fonction de la localisation visuelle. Pour bien relier les deux composantes d'une même source, nous devions donc renforcer la relation non plus spatiale mais temporelle et trouver des objets avec un synchronisme fort. Pour ce besoin de synchronisme, nous avons utilisé des vidéos avec des personnes parlant ou chantant.

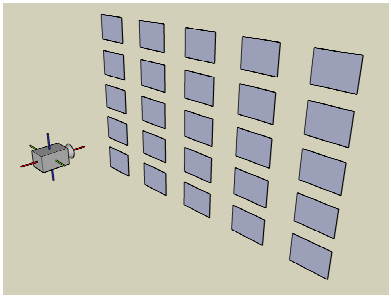


Figure 26. Schéma de disposition des éléments dans l'environnement créé, ici dans un cas abstrait de 25 vidéos. Le point de vue est représenté par la caméra positionnée au devant des vidéos.

Nous avons ainsi réalisé une scène où plusieurs vidéos, en nombre suffisant, sont réparties sur un plan devant la caméra, ou bien réparties sur plusieurs plans pour ajouter un effet de profondeur (Figure 26). De façon à pouvoir toujours identifier l'image de la vidéo et toujours garder le synchronisme visuel, les vidéos restent toujours tournées vers l'utilisateur. Par exemple, cette scène pourrait permettre à un utilisateur de voir et entendre

simultanément un ensemble d'extraits de films répartis devant lui et chercher quel est celui qu'il veut réellement regarder.

III. 4 - Interaction et navigation

Une fois les scènes décrites tant au niveau de l'image qu'au niveau du son (texture, taille des objets, fichier son associé, position dans l'espace, etc.), une question se pose encore : « comment l'utilisateur va-t-il interagir avec l'environnement virtuel pour trouver ce qu'il cherche dans cet espace frontal de la scène que nous avons décrite, en trois dimensions ? ». Cette question pose le problème à la fois de l'interface utilisée par l'utilisateur pour contrôler la scène, mais aussi celui de la méthode de navigation.

Les méthodes de navigation que nous proposons sont similaires à celles développées pour des interfaces uniquement visuelles mais combinent un rendu sonore, déformé ou non selon les techniques, cohérent avec le rendu visuel.

La méthode de navigation traditionnelle dans un espace navigable fonctionne par déplacement du point de vue et du point d'écoute de façon cohérente (point d'écoute et point de vue sont situés au même endroit), c'est-à-dire un déplacement de la caméra et du micro virtuel au sein de l'espace tridimensionnel. Nous avons ajouté à ce système la possibilité de gérer une lentille déformante audio-graphique que nous avons implémenté. L'utilisateur doit pouvoir gérer, en temps réel et en même temps, le point de vue/d'écoute et la lentille. L'ensemble des paramètres que l'on peut contrôler est alors en nombre important : position de l'ensemble caméra/micro, angle de prise de vue/de son (rotations de la caméra), position de la lentille sur l'écran, dimension de la lentille (rayon), focale de la lentille (coefficient de distorsion). Dans un premier temps, nous avons pu limiter le nombre de ces paramètres car la scène proposée était frontale. Nous avons simplifié les mouvements de la caméra. La méthode de navigation obtenue est de type Pan&Zoom ([2]) : la caméra se déplace de façon à ce que son axe reste orthogonal au plan de projection de la scène. Aucune rotation de la caméra n'est possible. Et comme, dans cette première phase de recherche, nous cherchons à garder une cohérence spatiale entre l'image et le son, le micro virtuel symbolisant le point d'écoute reste lui accroché à la caméra.

Nous avons d'abord mis en place un système de contrôle au clavier où l'utilisateur peut à la fois gérer le déplacement de la caméra et les paramètres de la lentille. Toutefois, en plus de ne pas être intuitive et même plutôt compliquée, cette interface ne permet pas de combiner la modification de deux paramètres simultanément.

C'est pourquoi un autre contrôleur a été proposé : la souris. Elle ne propose malheureusement que trois axes de contrôle et il faut donc choisir parmi les paramètres ceux qui peuvent être gérés soit par la position du curseur à l'écran, soit par la molette de la souris. Nous avons proposé deux possibilités. La première permet de gérer la caméra et convient au contrôle total de la méthode Pan&Zoom qui ne fait intervenir que trois paramètres : déplacements haut/bas et gauche/droite par la position du curseur (« *Panning* ») et mouvement avant/arrière (« *Zooming* ») avec la molette. La seconde possibilité gère la lentille déformante. Tout naturellement nous avons décidé d'associer le déplacement du curseur au déplacement de la lentille sur l'écran. La molette sert à gérer la distorsion, c'est à dire le niveau de *zoom*. La taille de la lentille est alors fixe. Aucune des deux méthodes n'est optimale et la question du contrôle fait partie des perspectives de recherches que nous allons mettre en place, en particulier pour la fin de ce stage.

IV. Expérimentation

Nous avons développé des outils pour la création de scènes virtuelles associant audio et visuel dans un espace en 3D. Pour faciliter l'exploration de ces environnements, nous avons mis en place deux systèmes de navigation. Le premier système, de type Pan&Zoom, permet de naviguer dans l'espace tridimensionnel en déplaçant une caméra et un microphone virtuels de façon cohérente. L'autre approche propose de déplacer une lentille grossissante type FishEye sur l'espace de projection en 2D. La phase du stage qui est présentée ici vise à tester et comparer ces deux méthodes de navigation.

IV. 1 - Les méthodes d'évaluation des techniques de navigation

IV. 1. a - La loi de Fitts

Pour évaluer les interfaces graphiques utilisateurs (GUI pour Graphical User Interfaces), et en particulier les interfaces zoomables, sur le plan maniabilité et efficacité, plusieurs méthodes expérimentales ont été proposées au cours de ces dernières années. Certaines sont basées sur des tâches spécifiques à un domaine, comme retrouver une ville sur une carte ou lire des documents textes. Malheureusement les résultats alors obtenus ne permettent souvent pas de conclure sur la validité des interfaces évaluées et soulèvent même parfois des contradictions. De plus, étant alors démontrés seulement sur un domaine relativement restreint, on ne peut que rarement les généraliser à d'autres champs d'applications. Des expérimentations plus contrôlées, faisant appel à ce que l'on pourrait appeler « tâches de laboratoire », aboutissent à des résultats plus significatifs. Ces évaluations s'appuient sur la **loi de Fitts**. Il s'agit d'un paradigme de pointage qui rend compte des capacités cognitives humaines de pointage et permet de mieux comprendre, voire de prédire, le temps mis pour trouver, viser et atteindre une cible (une icône ou un bouton cliquable par exemple). Si au départ cette loi s'appliquait à des cibles physiques dans le monde réel, elle a été généralisée pour plusieurs interfaces homme-machine, et en particulier pour les interfaces multi-échelles, comme celles à base de Pan&Zoom ([13]).

D'après la loi de Fitts, le temps T mis pour atteindre une cible, de largeur L et placée à une distance D, peut être estimée par la formule :

$$T = a + b \times \log_2(\frac{D}{L} + 1) \qquad \textbf{Équation 8}$$

où « $\log_2(\frac{D}{L} + 1)$ » est l'indice de difficulté (ID) et a et b sont deux coefficients empiriques.

Cette loi reprend ainsi le fait que pour une dimension de cible L fixée, si l'on recule la cible, il est plus difficile de l'atteindre (l'ID augmente) et le temps mis est plus long. De même , si l'on fixe la distance et que l'on place deux cibles côte à côte, l'une plus grande que l'autre, alors on mettra moins de temps à viser et atteindre la grande cible car ce sera plus facile (l'ID diminue) .Or l'intérêt principal des interfaces zoomables est de pouvoir jouer sur la dimension des objets (exemple *FishEye* ou *Bifocal Display*) ou sur la distance des objets (le Zoom dans la technique *Pan&Zoom*). Ces techniques permettent donc de modifier l'indice de difficulté. En se basant sur le coefficient de distorsion d'une lentille FishEye ou sur la distance parcourue par la caméra dans une vue Pan&Zoom, on peut estimer le temps minimum pour pointer une cible et prédire l'efficacité d'une méthode.

IV. 1. b - Tâches standardisées

Ainsi, une première tâche que l'on peut utiliser ([13] p. 20), pour valider ou comparer deux techniques de navigation, mesure le temps mis pour passer d'une cible à une autre. Deux cibles de même taille sont alors placées à l'écran face à l'utilisateur et celui-ci doit déplacer un curseur et cliquer sur chacune des deux cibles alternativement et ce le plus vite possible. On mesure alors le nombre de clics enregistrés en un temps préfixé. Cette tâche permet d'obtenir, expérimentalement, en faisant varier l'indice de difficulté (positions des cibles l'une par rapport à l'autre ou taille des cibles), une loi linéaire reliant le temps écoulé entre deux clics, c'est-à-dire T dans la loi de Fitts, et l'indice de difficulté. Cela vérifie donc la loi de Fitts.

Une autre tâche a donné lieu à une norme ISO (ISO 9241-9). Entre autres, Gutwin l'utilise dans ses expériences et la présente dans [14] : 24 cibles circulaires sont disposées sur un cercle. Le participant doit cliquer sur ces cibles, une par une, dans un ordre qu'il ne connaît pas. La prochaine cible à cliquer est désignée par une coloration verte et une croix rouge (Figure 27). Là encore, le but est de mesurer le temps mis par l'utilisateur pour atteindre la prochaine cible.

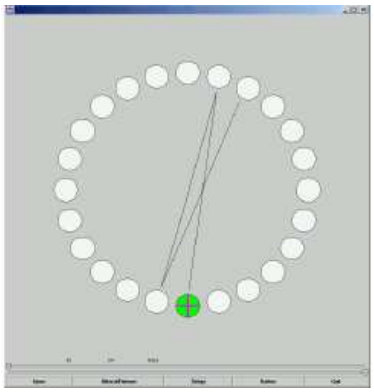


Figure 27. Vue de la répartition des cibles pour la tâche ISO 9241-9

IV. 1. c - Autre exemple de tâche

Parmi les tâches d'évaluation qui ont été proposées dans la littérature, l'une d'entre elles a retenu notre attention car elle peut s'appliquer aux scènes de vidéos que nous avons développées. Il s'agit d'une tâche présentée par Pietriga *et al.* dans ([22]).

Cette tâche consiste à trouver le plus vite possible une cible parmi un ensemble d'objets, en cherchant à faire le moins d'erreur possible. Ils présentent ainsi une grille de 3x3 objets bien alignés de façon à ce que l'utilisateur sache où sont les positions potentielles de la cible et n'ait pas en plus à chercher le lieu de la cible. En effet, son but est seulement d'identifier et de sélectionner l'objet qui correspond à la description de la cible. Les objets sont présentés par dessus un arrière plan carré et gris foncé. Les 9 objets sont des petits carrés sauf la cible qui est en fait un carré avec des coins arrondis. Pour voir apparaître l'arrondi il faut être à un niveau de zoom suffisant. Le participant observe ainsi chacun des objets en zoomant jusqu'à ce qu'il ait trouvé la cible.

IV. 2 - Protocole expérimental

IV. 2. a - Les systèmes comparés et leur modes d'interactions

Notre but est d'évaluer le système de lentille déformante audio-graphique que nous avons mis en place durant ce stage. Il s'agit d'une lentille déformante du type *FishEye* (FL pour *FishEye Lens*), que l'on peut déplacer sur la surface de représentation de façon à grossir certains objets tout en gardant le même contexte d'ensemble. La distorsion présentée n'est plus uniquement visuelle mais également sonore. De sorte que le son des objets situés dans la zone définie par la lentille est également grossi (niveau sonore plus élevé) mais également espacé du centre de la lentille. L'utilisateur se sert d'une souris, soit pour déplacer la lentille (le curseur est remplacé par le centre de la lentille), soit pour changer le coefficient de distorsion *d* (niveau de focus) par la molette. La distorsion est alors équivalente entre l'audio et le graphique.

Pour évaluer cette technique, nous avons cherché à le comparer à un autre mode d'exploration de l'environnement que nous avions à disposition et qui avait été implémenté au cours de précédentes recherches sur les espaces navigables. Il s'agit d'une méthode Pan&Zoom (PZ) traditionnelle où une caméra est placée orthogonalement au dessus de la surface de présentation des objets, et se déplace selon les axes gauche/droite et bas/haut (*pan*), mais peut aussi se rapprocher ou s'éloigner de ce plan dans une mouvement d'axe avant/arrière (*zoom*) permettant de changer le niveau de détails des objets. Notre méthode, étendue au multimédia permet de gérer de façon cohérente et simultanée le point de vue (caméra) et le point d'écoute car le micro virtuel est attaché à la caméra. Avec cette méthode, l'utilisateur peut contrôler le déplacement gauche droite et le déplacement bas-haut en se servant d'une souris. Le déplacement avant/arrière qui correspond donc au niveau de détail est géré par la molette de la souris.

IV. 2. b - Choix de la tâche

Dans un premier temps, nous avons cherché à reprendre notre scène frontale constituée de vidéos pour ses intérêts de synchronisme image-son. Comme dans l'expérience de Pietriga *et al.* présentée en IV. 1. c -, notre idée est de demander aux participants de retrouver, le plus rapidement possible et en limitant les erreurs, une vidéo-cible parmi toutes les vidéos présentées. Au départ, les objets sont tous visibles et alignés sur une grille pour que le participant n'ait pas en plus à chercher la position éventuelle de la cible. Mais pour être sûr que le participant utilise les techniques de zoom, et que ce sont donc bien les techniques d'interaction qui sont comparées, les objets sont présentés au départ de façon à ce qu'on ne distingue pas le contenu visuel (taille réduite des vidéos) et que l'on ne puisse rien entendre (niveau sonore trop faible).

De plus, le nombre de vidéos présentées simultanément est restreint car si l'on ne dispose pas les vidéos en fonction de critères prédéfinis comme on le fait pour classer les données dans des bases (indexation), il peut être très long de parcourir l'ensemble des vidéos avant de trouver le bon élément. Cette limitation du nombre de vidéo est appliquée de la même manière pour les deux techniques comparées, donc la validité écologique de la tâche n'est pas diminuée.

Les différentes vidéos utilisées sont des extraits du concours de l'Eurovision. Chaque

extrait présente donc un chanteur ou un groupe différent. Cela permet d'avoir des vidéos avec un fort synchronisme entre l'image et le son et présentant toutes une identité visuelle et musicale forte. De plus la qualité de rendu entre chacune des vidéos est identique. Les vidéos durent toutes 10 secondes et sont jouées en boucle.

IV. 2. c - Hypothèses/prédictions

Nous cherchons à évaluer les améliorations apportées par le nouvel outil de navigation proposé. Le but de cette expérience est donc avant tout de montrer que la technique FL par distorsion audio et visuelle de l'espace est à la fois plus rapide, plus agréable et plus performante que la technique PZ de déplacement de la caméra.

Cette hypothèse se base sur le fait que l'exploration nécessite moins d'opérations différentes de la part de l'utilisateur avec une lentille qu'avec la technique Pan&Zoom. En effet, avec la méthode PZ, pour passer d'un objet au suivant, l'utilisateur doit alors faire un zoom arrière pour retrouver une vue d'ensemble et repérer le prochain objet, puis déplacer la vue pour la centrer sur cet objet et enfin faire un zoom avant vers le nouvel objet pour retrouver le niveau de détail souhaité. En revanche, avec une lentille, une fois le niveau de zoom atteint pour l'objet que l'on examine, il suffit de déplacer la lentille pour atteindre l'objet suivant directement avec le niveau de détail souhaité.

Cette expérience devrait également vérifier que le nombre d'objets présentés en même temps, et incarnant le rôle de cibles potentielles, va influencer le temps moyen nécessaire pour trouver la vidéo cible parmi toutes celles proposées. En effet, plus le nombre de vidéos, parmi lesquelles l'utilisateur doit trouver la cible, augmente et plus le nombre de vidéos que l'utilisateur est susceptible d'essayer avant de trouver la bonne augmente, entraînant une augmentation de la durée prise pour retrouver la vidéo cible. Nous tenons alors compte de la chance et malchance potentielles des participants qui peuvent très bien trouver la cible "du premier coup", comme au contraire ne trouver enfin la cible qu'après avoir vérifié que toutes les autres vidéos n'étaient pas celles recherchées. En moyenne, l'utilisateur devra donc tester la moitié des vidéos avant de trouver la vidéo recherchée.

IV. 2. d - Les variables

Parmi les différents paramètres qui peuvent varier au cours de cette expérience, nous distinguons les variables **indépendantes,** que l'on peut contrôler, et les variables **dépendantes,** qui sont celles que l'on mesure.

Variables indépendantes

Les variables indépendantes sont donc dans notre cas :

- La technique utilisée : PZ (déplacement de l'ensemble caméra/micro) ou FL (déplacement d'une lentille déformante audio-visuelle).
- le nombre de vidéos présentées : nous avons décidé de comparer dans un premier temps des scènes avec 9 et 25 vidéos.
- le niveau d'expertise des sujets, notamment en matière de son 3D, d'interactions

homme-machine et d'exploration de données.

- Position de l'objet recherché au début de la vidéo. Pour gérer ce point, nous avons décidé de mettre en place un système de tirage aléatoire et de répartition aléatoire des différentes vidéos sur la grille.
- La taille des objets et le niveau sonore des vidéos avant le changement d'échelle. Dans cette première expérience, nous avions fixé ces paramètres pour qu'on ne puisse ni entendre ni voir les vidéos au départ.

Variables dépendantes

Les paramètres que l'on va mesurer sont quant à eux le nombre d'erreurs, le temps écoulé pour trouver et sélectionner la vidéo recherchée, ainsi des commentaires subjectifs des sujets comme leur préférence entre les deux méthodes pour réaliser la tâche où le système avec lequel ils pensent avoir été plus rapide.

IV. 2. e - Contrebalancement

Pour éviter qu'un biais dû à l'apprentissage n'apparaisse entre le passage d'une méthode à une autre, nous avons fait deux groupes de participants. Le premier groupe débute l'expérience en utilisant la technique 1 (PZ) puis continue par la technique 2 (FL). L'autre groupe procède à l'inverse (technique FL puis PZ). En revanche l'ordre de passage pour le nombre de vidéos n'a pas été contrebalancé. Nous sommes en effet partis du principe que la tâche devenait suffisamment complexe avec 25 vidéos sources, et donc que l'apprentissage ne serait pas inutile.

Nous avons alors comme ordre de passage pour chaque groupe :
Groupe 1 : technique PZ et 9 vidéos, technique FL et 9 vidéos, technique PZ et 25 vidéos et technique FL et 25 vidéos.
Groupe 2 : technique FL et 9 vidéos, technique PZ et 9 vidéos, technique FL et 25 vidéos et technique PZ et 25 vidéos.

De plus, nous avons 4 conditions (2 techniques * 2 nombre de vidéos) et, pour chaque condition nous avons mise en place, 8 essais. Ainsi chaque participant réalise 2*2*8 = 32 essais (ce qui correspond à environ 50 secondes/essai soit environ 27 minutes).

IV. 2. f - Scénario

L'expérience est séparée en 3 parties :
une phase d'apprentissage où le participant peut essayer chacune des deux techniques, une phase de test puis une phase de questionnaire.

Les deux premières phases sont constituées de plusieurs séries "d'essais". A chaque essai, une nouvelle vidéo cible, tirée aléatoirement, est présentée au participant à qui il est demandé d'observer attentivement. Puis le sujet déclenche lui même la partie recherche de l'essai en appuyant sur une touche du clavier. Il pourra revenir par la suite à tout moment à la vidéo cible s'il ne s'en souvient plus. L'exploration se fait alors à l'aide de la souris avec l'une des deux méthodes présentées auparavant. Quand l'utilisateur est suffisamment près ou a zoomé suffisamment sur une vidéo, elle devient sélectionnée (entourée d'un bord

rouge). Si l'utilisateur pense que c'est la bonne, il peut valider. Il doit pour cela presser une autre touche du clavier. Puis une nouvelle vidéo cible est tirée au sort et on recommence. Le storyboard d'un essai est présenté en Figure 28.

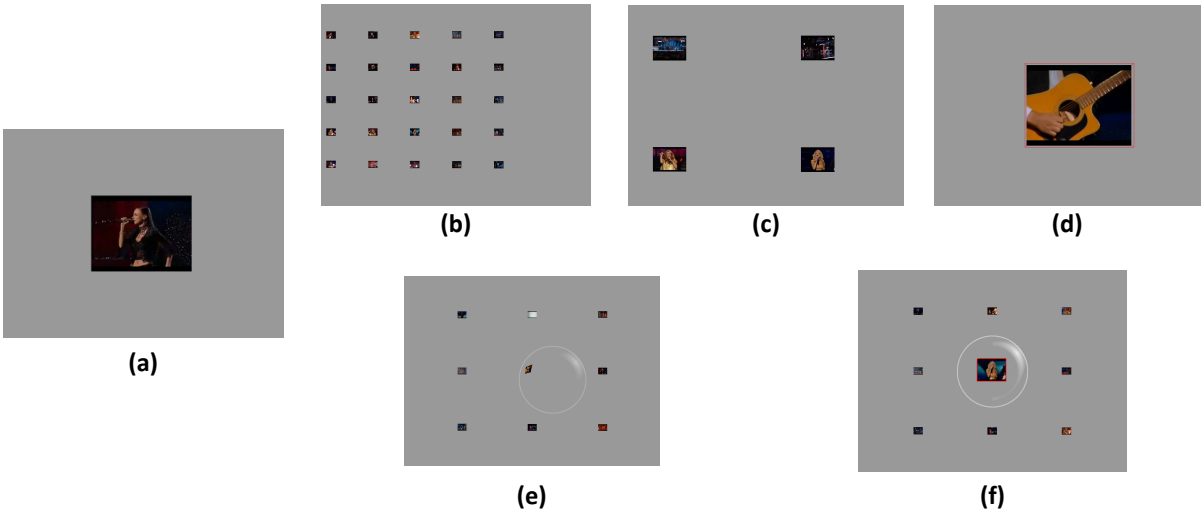


Figure 28. Storyboard d'un essai avec chacune des deux méthodes d'exploration comparées. a) lecture d'une vidéo cible : absence de distorsion. b) au départ de l'exploration pour 25 vidéos et la méthode PZ. puis changement d'échelle et déplacement (c) et enfin sélection et validation (d). e) début de changement d'échelle avec la méthode FL et 9 sources puis sélection et validation (f).

Pour la phase d'apprentissage, les séries comprennent 3 essais uniquement, de façon à ce que l'utilisateur se familiarise juste avec le scénario d'un essai : lecture de la vidéo cible, lancement de la partie exploration, exploration, sélection, validation. Pour la phase de test, c'est 4 séries (une série par condition) de 8 essais que le participant devra effectuer. Les conditions sont bien séparées de façon à limiter le temps d'adaptation nécessaire entre chaque essai. A la fin du test, le participant est invité à répondre à un questionnaire qui lui permet de préciser la catégorie de sujet à laquelle il appartient (âge, sexe, questions sur l'expertise), mais aussi à donner son avis sur les techniques utilisées. En tout, l'évaluation dure environ 40 minutes par participant.

IV. 3 - Analyse

IV. 3. a - Les participants

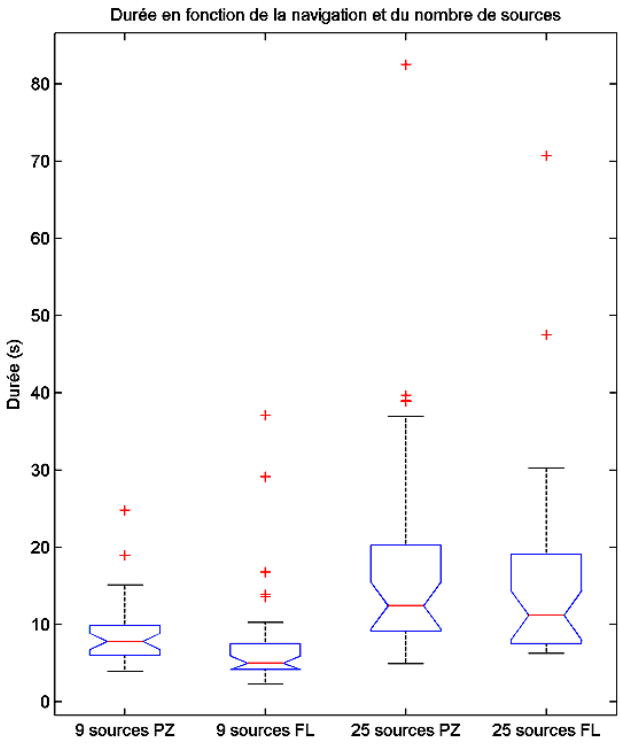
Pour le moment seules 4 personnes ont pu participer à cette expérience. Nous espérons faire participer un minimum de 12 personnes par la suite. Pour le moment les 4 participants sont de sexe masculin, âgés de 21 à 24 ans, et sont tous familiarisés avec la pratique informatique et en particulier avec l'utilisation de la souris.

IV. 3. b - Comparaison objective : analyse des mesures

Dans un premier temps nous avons vérifié le nombre d'erreurs effectuées au cours des 32 essais. Sur les 4 sujets, aucune erreur n'a été détectée. Tous ont donc trouvé la bonne vidéo à chaque fois, et c'est seulement sur le temps mis pour aboutir à cela que nous pouvons nous appuyer.

En se basant sur la distribution des résultats obtenus (Figure 29 b), on peut voir qu'ils ne sont pas vraiment exploitables pour un nombre de 25 vidéos car l'étendue des résultats est trop grande. On peut expliquer cela par le nombre d'essai trop restreint par personne (8 essais) alors que le participant a le choix entre 25 sources. Cela est confirmé par la Figure 30 (page suivante) qui présente la distribution des résultats obtenus en fonction du nombre de vidéos pour chacun des 4 sujets. La variabilité de temps mis pour trouver la cible augmente considérablement quand on a 25 sources. Avant de faire passer d'autres sujets, il sera nécessaire de modifier légèrement le protocole expérimental pour que chaque participant passe plus d'essais par condition.

Condition	Moyenne (s)	Ecart-type
PZ- 9 sources	8.9	3.7
FL- 9 sources	7.9	6.4
PZ-25 sources	17.8	14.8
FL-25 sources	16.0	11.4



(a) **(b)**

Figure 29. a) Durées moyennes et écarts-type par essai en fonction du nombre d'objets présentés simultanément et de la technique employée. b) Distribution des résultats de durée : la barre rouge représente la médiane, la boîte bleue contient 50% des résultats, les barres transversales (« moustaches ») 90%, les croix rouges représentent les valeurs les plus éloignées de la médiane.

Cependant, les résultats obtenus affichent déjà nettement des tendances.

Aussi voit-on déjà clairement apparaître une augmentation du temps de recherche lorsque le nombre de sources augmente (Figure 30), ce qui est valable pour tous les participants. Or c'était bien les résultats escomptés. Toutefois nous ne pourrons valider cela qu'une fois que l'étendue des résultats sera moins importante.

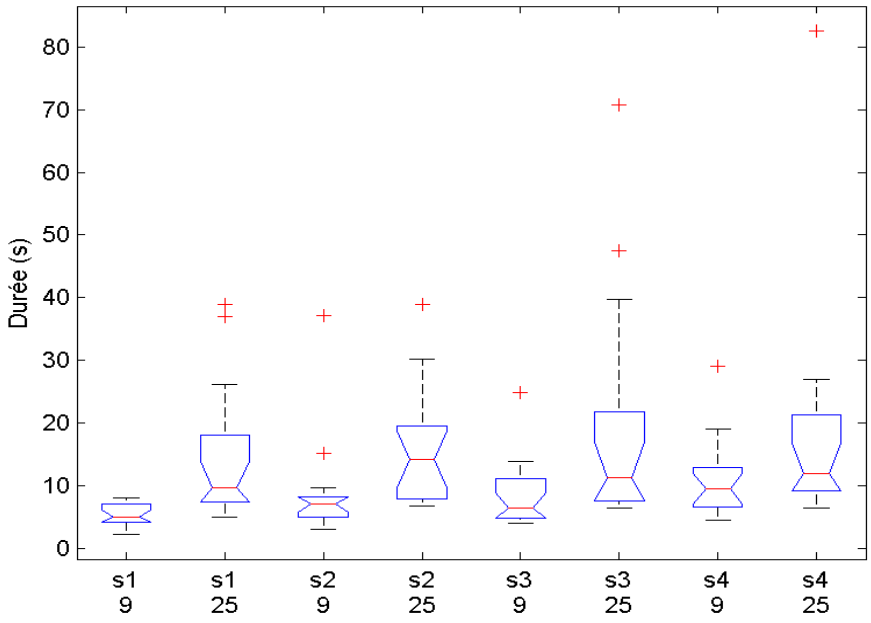


Figure 30. Comparaison de durées en fonction du nombre de vidéos présentées simultanément pour chacune des 4 sujets.

De la même manière, le temps mis en moyenne pour trouver la vidéo montre que la technique par manipulation de la lentille est plus efficace, car plus rapide, que la technique par déplacement de la caméra (Figure 31 a), et ce quelque soit le nombre de vidéos présentées. Et l'on remarque que si les résultats sont très étendus lorsque l'on considère l'ensemble des participants, en revanche, l'observation tient pour tous les participants : la vidéo est trouvée et validée plus vite avec la méthode *FishEye* (Figure 31 b). Là encore nous obtenons les résultats attendus.

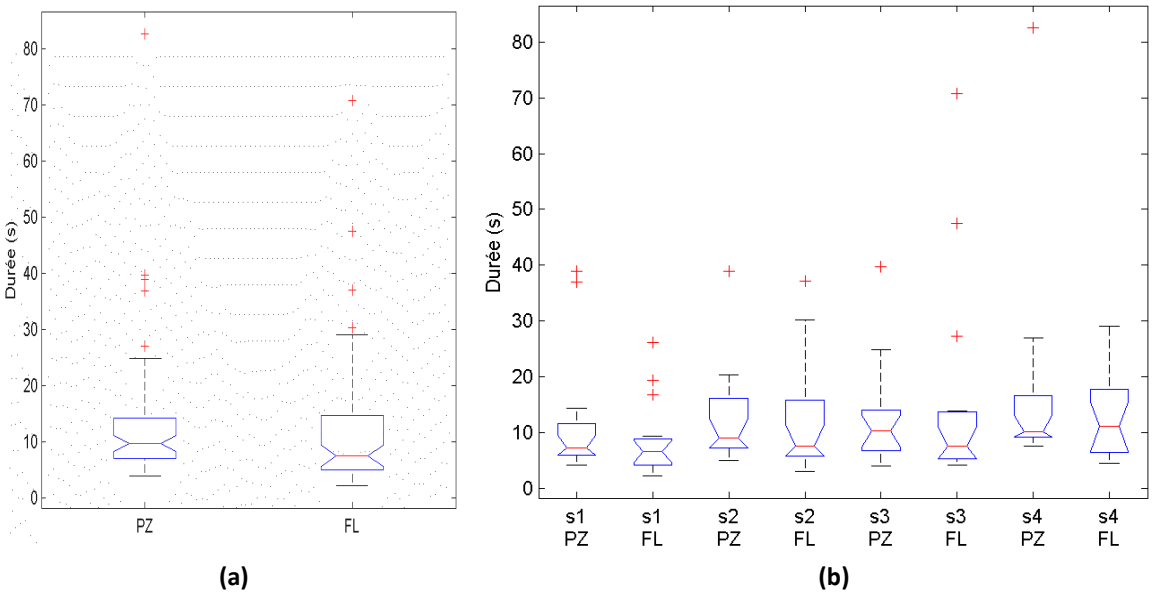


Figure 31. a) Comparaison de la durée moyenne pour chacune des deux méthodes, pour tous les sujets et pour les cas de 9 et 25 vidéos confondus. b) Comparaison des deux méthodes pour chacun des sujets, les cas de 9 et 25 vidéos confondus.

IV. 3. c - Analyse subjective : le questionnaire

Toutes les personnes interrogées à la suite du test ont préféré la technique par déplacement de la lentille et se sont trouvés plus rapides. Dans les raisons citées par les participants pour expliquer leur préférence, celle qui revient le plus (3 fois sur 4) est que cette méthode permet une vision de l'ensemble contrairement à l'autre méthode. Ainsi un des participants justifie par exemple son choix par « avec la lentille, on voit à chaque instant l'ensemble de l'espace. Avec le mouvement de la caméra, on se promène au hasard jusqu'à trouver la bonne vidéo ».

Un des participants souligne toutefois que pour lui c'est une question de contrôleur : la souris est parfaitement adaptée selon lui à déplacer une lentille sur un écran mais ce n'est pas intuitif pour déplacer la caméra.

Tous les sujets ont également souligné, lors de commentaires libres, qu'il étant gênant de ne pas pouvoir entendre tout de suite les sons au début. Ils reprochent, d'une certaine manière, au système d'obliger à zoomer, et ce pour les deux techniques, pour entendre les sons. On peut comprendre cela comme une volonté d'avoir un contexte sonore dès le début afin de pouvoir faire un premier repérage. Cela est très encourageant pour la suite de nos recherches car cela semble montrer que les distorsions focus+contexte au niveau audio seront appréciées dans le cadre d'environnements plus réalistes. De plus, cela confirme la nécessité de proposer un rendu sonore dans le cas de base de données audio-visuelles.

IV. 4 - Critiques du protocole expérimental et mise en place de la version suivante

En plus des problèmes liés au nombre d'essais que nous avons relevés en analysant les résultats du test, nous avons également conscience de certains défauts de cette évaluation que nous pourrons modifier pour une expérience à réaliser dans un futur proche.

IV. 4. a - Problématique de la tâche et du choix des vidéos

La tâche qui était à effectuer dans le protocole expérimental précédent faisait intervenir des procédés de reconnaissance de visage, de voix, de musique et d'image animée. Or nous n'avions pas de connaissances préalables en matière de cognition et nous avons donc fait abstraction de cette partie. Or nous ne pouvons pas être sûre que les vidéos que nous avons choisis n'influencent pas les résultats. Les analyses effectuées ne sont valables qu'avec ces vidéos liées à de la musique. La tâche, trop spécifique, ne permet donc pas de généraliser. Il faudrait recommencer avec d'autres genres de vidéos, avec de la parole par exemple ou avec des objets plutôt qu'avec des humains.

Un autre jeu de vidéos auraient dû être prévu à l'avance pour contrebalancer cet effet, mais cela aurait augmenté encore la durée de l'expérience.

IV. 4. b - Choix d'une tâche trop compliquée faisant intervenir plusieurs modalités en même temps

La distinction entre les modalités auditive et visuelle a été prise en compte mais n'a

malheureusement pas été intégrée à cette expérience qui aurait rendue la durée du test extrêmement longue pour les participants. Il est cependant nécessaire de réaliser une autre expérience prochainement afin de rendre compte de l'apport de chaque modalité au système. Pour cela, nous allons comparer non plus des techniques de navigation, mais des ensembles de rendus. Ainsi, nous pensons garder le même type de protocole expérimental basé sur une recherche de vidéo, mais en comparant trois modalités d'exploration différentes:

- distorsion audio-visuelle, comme c'est le cas dans le protocole expérimental déjà mis en place,
- distorsion audio seule : l'aspect graphique n'est pas modifié mais on peut zoomer auditivement sur les objets. Cependant pour ne pas perdre l'utilisateur dans la masse de données alors sonores, nous devrons marquer d'un signe visuel sa position ou celle de la lentille.
- distorsion graphique seule : le rendu audio reste inchangé.

Nous avons pensé également à comparer les résultats du test déjà mis en place avec ceux d' une expérience similaire mais sans rendu audio. L'hypothèse est alors qu'en présence du rendu audio les sujets seront plus rapides pour trouver l'information recherchée.

V. Conclusion

Au cours de ce stage, plusieurs outils pour l'exploration d'espaces 3D audiovisuels ont été présentés. En particulier, en s'inspirant des méthodes visuelles d'interaction pour la visualisation d'informations, nous avons proposé une technique de distorsion de type lentille grossissante *FishEye* agissant à la fois sur les modalités auditives et visuelles. Une évaluation, basée sur les techniques d'évaluation des interfaces zoomables, a été mise en place, et les premiers résultats ont permis de montrer un gain de temps, ainsi qu'une amélioration de confort, grâce à cette approche par comparaison à une technique habituelle de navigation dans des espaces 3D.

Le stage n'est cependant pas encore terminé et d'autres éléments devraient être mis en place durant le dernier mois. De plus, le champ d'étude sur les rendus non réalistes est large et nous aimerions exploiter, au cours de futures études, plusieurs pistes de recherche auxquelles nous avons déjà pensées.

VI. Perspectives

VI. 1 - A court terme

VI. 1. a - Préparation d'autres scénarii audiovisuels incluant la profondeur

La mise en scène développée durant le stage visait à répartir plusieurs vidéos sur un plan faisant face à l'utilisateur. Cet environnement pouvait s'apparenter à une représentation uniquement en 2D. Pour avoir une réelle extension à des espaces tridimensionnels, nous avons pensé mettre en place d'autres environnements virtuels faisant intervenir cette fois la notion de profondeur. Une idée de scène consiste en la création d'un monde constitué d'une foule où l'utilisateur aurait une vue légèrement du dessus, tel un dieu qui surveillerait tous ces gens. Nous pensons alors à la possibilité de faire en sorte que les avatars alors présents dans le monde virtuel parlent tous, en même temps et de façon très désordonnée. L'utilisateur ne peut alors pas comprendre les paroles (effet cocktail party[2]) et pour écouter une personne en particulier doit utiliser les outils de navigation et de distorsion audio mis en place. Contrairement à la scène utilisée durant le stage, sur la recherche de vidéos, il s'agit d'un environnement réaliste.

VI. 1. b - Mise en place d'une méthode de navigation permettant de gérer à la fois la caméra et la lentille

Actuellement, le contrôle mis en place est limité. La souris est en effet un contrôleur ne présentant que peu de degrés de liberté et nous ne pouvons faire varier que plusieurs paramètres à la fois. De plus la combinaison de la souris et du clavier ne résout pas le problème de l'interface car le contrôle, non intuitif, devient difficile. Pour gérer à la fois les paramètres de la caméra et ceux de la lentille déformante, nous proposons de mettre en place un autre dispositif de contrôle basé sur la captation du geste. Ainsi, la combinaison entre une wiimote, pour la capture des rotations et des déplacements gauche/droite de la caméra, et le joypad présent sur cette manette devraient permettre de contrôler à la fois le point de vue/d'écoute et les paramètres de la lentille.

VI. 2 - A plus long terme

VI. 2. a - Installation d'un système immersif de grande ampleur

L'avantage du dispositif mis en place, combinant rendu visuel en trois dimensions et rendu sonore spatialisé, est qu'il permet de créer des environnements immersifs.

Pour la partie visuelle, nous comptons renforcer la profondeur de l'image par un rendu stéréoscopique jouant sur la parallaxe entre deux rendus visuels de la scène légèrement décalés, et superposés par projection sur un même écran.

Le rendu sonore est déjà en trois dimensions. Cependant, nous comptons développer un dispositif pour avoir un espace d'immersion plus grand : tout d'abord le rendu audio ne

[2] http://fr.wikipedia.org/wiki/Effet_cocktail_party

se fera plus au casque pour éviter un dispositif trop intrusif. Nous utiliserons alors un système de rendu par Wave Field Synthesis à l'avant ce qui permettra d'y obtenir un rendu précis. A l'arrière nous pourrons également placer des enceintes (non gênantes visuellement car absentes du champ visuel) et obtenir un rendu ambisonique. La combinaison de ces deux modes de spatialisation sonore devrait permettre une sensation d'espace sonore englobant, avec une grande précision de localisation permettant alors des distorsions fines.

VI. 2. b - Mise en place d'autres types d'interfaces zoomables dans des environnements multimédias

Nos avons développé durant ce stage une technique de distorsion de l'espace par une lentille grossissante. D'autres techniques de distorsion focus+contexte peuvent être exploitées. En particulier une techniques mise en œuvre dans le cadre d'interfaces zoomables consiste à superposer en transparence le focus et le contexte présentant simultanément les deux vues et ce sans masquage ni déformations ([19]). Nous pourrions reprendre cette idée en superposant deux points de vue sonores. Une des applications pourrait se trouver alors dans des environnements plus réalistes, comme une ville où l'on pourrait entendre simultanément l'ambiance sonore de l'extérieur et de l'intérieur des bâtiments, avant de se décider à rentrer définitivement dedans.

VI. 2. c - Extension de la notion de stylisation à l'audio

Pour répondre à des contraintes de temps, nous avons limité notre étude durant ce stage aux techniques de rendus non réalistes jouant sur la distorsion de l'espace. D'autres techniques non photoréalistes pourront être reprises et étendues à l'audio, en particulier les techniques par imitation de styles. Des traitements du son pourront alors être exploités, comme par exemple la granulation ([21]) qui peut se rattacher à la notion de texture et aux méthodes dites pointillistes du non photoréalisme. D'autres types de traitement pour l'audio seront également recherchés, notamment pour permettre une *cartoonification* sonore.

Les études futures passeront alors, comme pour les techniques non photoréalistes, par l'étude des points saillants des scènes auditives afin de pouvoir simplifier, comme en non photoréalisme, les scènes sonores et multimédias ([3],[9]). En vision il s'agit du repérage des contrastes et des couleurs afin de déterminer par exemple le contour et les silhouettes.

En jouant sur le changement de texture ou de couleur (le timbre sonore est souvent appelée « couleur » en musique), les traitement appliqués au son pourront également permettre, en superposant deux rendus sonores différents, d'obtenir d'autres types de lentilles magiques que les lentilles grossissantes.

Bibliographie

[1] Bouchara, T. « *Le "SceneModeler" : des outils pour la modélisation de contenus multimédias interactifs spatialisés* ». 13ème Journées d'Informatique Musicale (JIM'08), GMEA-AFIM, Albi, mars 2008.

[2] Bourgeois, F., Guiard, Y., Beaudouin-Lafon, M. « *Pan-Zoom Coordination in MultiScale Pointing* ». Conference on Human Factors in Computing Systems (CHI '01), 2001.

[3] Bregman, A. Auditory Scene Analysis : The Perceptual Organization of Sound. Cambrige MA, MIT Press, 1990.

[4] Coleman, P. et Singh, K. « *RYAN : Rendering Your Animation Nonlinearly projected* ». Proc. of the Third Symposium on Non-Photorealistic Animation and Rendering, NPAR, Annecy, France, juin 2004.

[5] Cotterell, P. S. "On the Theory of the Second-Order Soundfield Microphone", PhD Thesis, University of Reading, 2002

[6] Daniel, J. « Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia ». Thèse de doctorat, université Paris 6, 2000.

[7] Daniel, J., Nicol, R. et Moreau, S. "*Further Investigations of Higher Order Ambisonics and Wavefield Synthesis for Holophonic Sound Imaging*". 114th AES Convention, Amsterdam, mars 2003.

[8] DeCarlo, D. et Santella, A. "*Stylization and abstraction of photographs*". Proc. 29th annual conference on Computer graphics and interactive techniques, SIGGRAPH '02, SESSION: Painting and non-photorealistic graphics, San Antonio, Texas, 2002

[9] Gallo, E. « Restitution Sonore Hiérarchique et Perceptive d'Environnements Virtuels Multi-Modaux», Thèse de doctorat en Informatique, 2006.

[10] Gerzon, M. A. "*Ambisonics in Multichannel Broadcasting and Video*". J. Audio Eng. Soc., vol. 33 n°11, p. 859-871, Novembre 1985.

[11] Gerzon, M. A. "*Ambisonics Decoders for HDTV*", 92nd AES Convention, preprint 3345, march 1992.

[12] Green, S. « Introduction to Non-Photorealistic Rendering », SIGGRAPH, course on NPR, Chap. 2, 1999.

[13] Guiard, Y., Beaudouin-Lafon, M. « *Target Acquisition in Multiscale Electronic Worlds* ». International Journal of Human Computer Studies (IJHCS), vol. 61, n°6, Décembre 2004.

[14] Gutwin, C. "*Improving Focus Targeting in Interactive FishEye Views*". Proc. SIGCHI conference on Human factors in computing systems: Changing our world, changing ourselves, SESSION: Focue and Context? USA, 2002.

[15] Hascoët, M. et Beaudoin-Lafon, M. « *Visualisation interactive d'information* ». Revue I3, vol 1 n°1, 2001.

[16] Jacquemin, C., Folch, H., Garcia, K. et Nugier, S. "*Visualisation interactive d'espaces documentaires*". Revue I3, Information-Interaction-Intelligence, vol. 5, n°1.

[17] Lansdown, J. et Shofield, S. « *Expressive Rendering : A Review of Nonphotorealistic Techniques*». In IEEE Compter Graphics and Applications, vol. 15, n°3, 1995

[18] Lecolinet, E . et Nguyen, D. « *Représentation focus+contexte de listes hierarchiques zoomables*». Actes des 18e conférences francophones sur l'Interaction Homme-Machine (IHM'06), ACM Press, avril 2006.

[19] Lecolinet, E. et Pook, S. "*Interfaces zoomables et « Control menus » : Techniques focus+contexte pour la navigation interactive dans les bases de données*". Revue Les Cahiers du numérique. Vol.3, pp. 191-210. Hermès, Paris, Dec. 2002.

[20] Leung, Y. K. et Apperley, M. D. "*A review and taxonomy of distortion-oriented presentation techniques*". ACM Transactions on Computer-Human Interaction (TOCHI), Juin 1994.

[21] Pekonen, J. "Computationally Efficient Music Synthesis, Method and Sound Design". Chap. 2.2 Granular Synthesis, Master Thesis, 2007.

[22] Pietriga, E., Appert, C. M. Beaudouin-Lafon "*Pointing and Beyond: an Operationalization and Preliminary Evaluation of Multi-scale Searching*". Proc. ACM Conference on Human Factors in Computing Systems (CHI '07),2007.

[23] Pulkki, V. « Virtual sound source positioning using Vector Base Amplitude Panning". JAES, vol. 45, n°6, june 1997.

[24] Seitz, S.M. et Kim, J. "*Multiperspective imaging*". Computer Graphics and Applications, IEEE, Vol. 23, n° 6, Nov.-Dec.

[25] Smith, K. "Perspective : multiple perspective, nonlinear projection", presentation on SIGGRAPH, 2005.

[26] Stone, M. C., Fishkin, K., et Bier, E. A. "*The movable filter as a user interface tool"*. Proc. SIGCHI Conference on Human Factors in Computing Systems: Celebrating interdependence. CHI'94, ACM, Boston, Massachusetts, United States, Avril 1994.

[27] Sutcliffe, A. G., Kurniawan, S. et Shin, J-E. « *A method and advisor tool for multimedia user interface design*». International Journal of Human-Computer Studies, vol. 64, 2006.

[28] Viega, J., Conway, M. J., Williams, G., et Pausch, R. "*3D magic lenses*". Proc. 9th Annual ACM Symposium on User interface Software and Technology ,Seattle, Washington, United States, Novembre 1996.

ANNEXE 1 : Compléments sur la décomposition en harmoniques sphériques

Dans un système de coordonnées sphériques, la pression acoustique peut s'exprimer en une décomposition de Fourier-Bessel:

$$p(kr,\theta,\varphi) = \sum_{m=0}^{\infty} i^m J_m(kr) \sum_{n=0}^{m} B_{mn}^{\sigma} Y_{mn}^{\sigma}(\theta,\varphi) \qquad \textbf{Équation 9}$$

Où : les Y_{mn}^{σ} sont les harmoniques sphériques telles que présentées en Figure 33,

θ est l'azimuth (angle dans le plan horizontal),

φ le site (elevation an anglais, angle dans le plan vertical),

σ = ± 1, $i = \sqrt{-1}$,

mn est l'ordre de l'harmonique sphérique, $n \leq m$,

Les $J_m(kr)$ sont les fonctions de Bessel sphériques de premières espèces (Figure 32)

k est le nombre d'onde,

r est la distance entre l'origine du système de coordonnée et le point de mesure de la pression.

Chaque composante B_{mn}^{σ} est le canal ambisonique obtenu par projection orthogonale de la pression *p* sur l'harmonique sphérique correspondante.

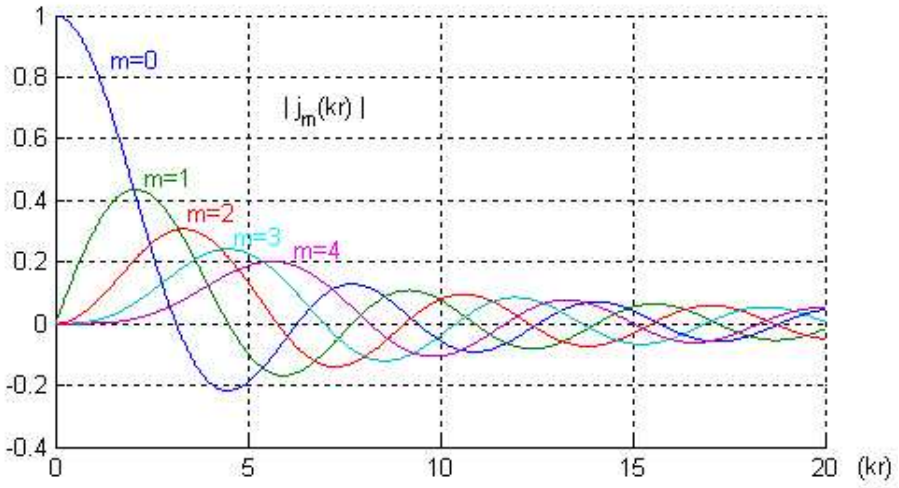


Figure 32. Fonctions de Bessel sphériques de première espèce $J_m(kr)$

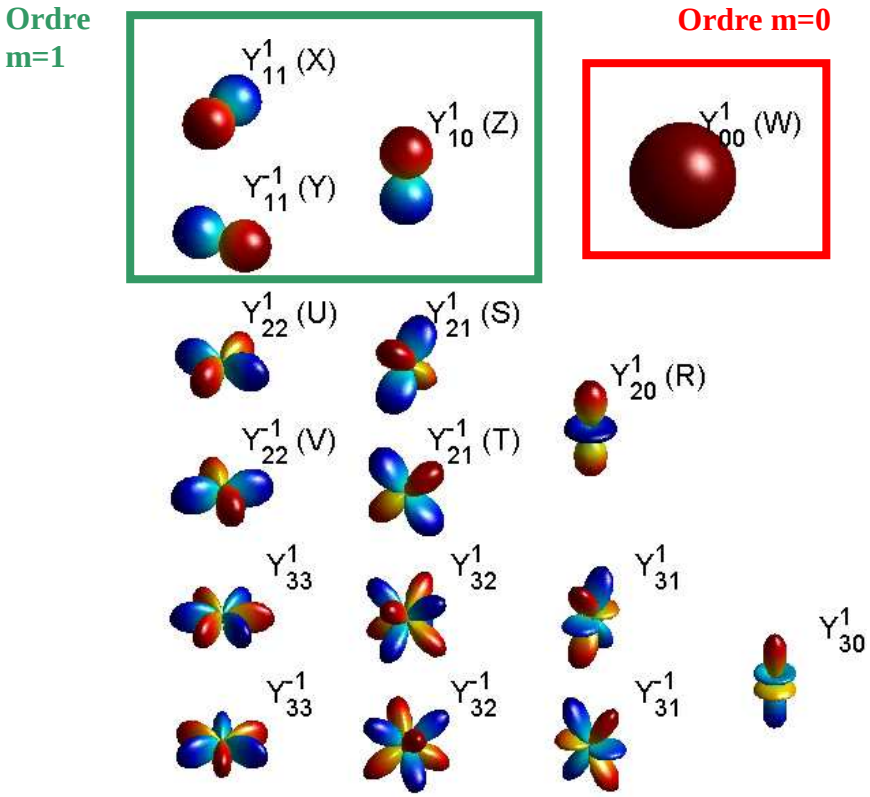


Figure 33. Vue 3D des harmoniques sphériques Y_{mn}^{σ} . En rouge, les valeurs positives, en bleu les valeurs positives des harmoniques sphériques. On a ainsi (2m+1) composantes, avec 2 composantes horizontales (n=m) par ordre m≥1.

ANNEXE 2 : Le décodage ambisonique

Rappel : *Il s'agit de reconstruire dans une zone d'écoute le champ acoustique précédemment encodé en fonction du système de diffusion ou d'écoute utilisé. On va en fait reconstruire le champ ambisonique B' au centre du système de restitution. Il s'agit donc de récupérer les signaux HOA (B_{mn}^{σ}), et d'obtenir à partir d'un matriçage de ces composantes un signal que chaque haut-parleur (HP) doit émettre. Il est important de souligner que le nombre de HPs nécessaires pour obtenir une bonne restitution (homogène sur tout l'espace) dépend de l'ordre M d'encodage. Il est fortement recommandé d'utiliser au moins L=2M+2 haut-parleurs.*

On considère pour cette étape que les HPs sont situés suffisamment loin de l'auditeur (au centre) et émettent donc des ondes planes. Chaque HP est repéré par sa position (θ_i,δ_i). Les signaux S_HP_i émis par chaque HP participent au champ ambisonique B' au centre O du système de diffusion de sorte que l'on ait :

$$B' = C.S \qquad \textbf{Équation 10}$$

avec

$$c_i = \begin{pmatrix} Y_{00}^{+1}(\theta_i,\delta_i) \\ Y_{11}^{+1}(\theta_i,\delta_i) \\ Y_{11}^{-1}(\theta_i,\delta_i) \\ \cdots \\ Y_{mn}^{\sigma}(\theta_i,\delta_i) \\ \cdots \end{pmatrix} \quad C = \begin{pmatrix} c_1 & \cdots & c_i & \cdots & c_L \end{pmatrix} \quad B' = \begin{pmatrix} B'^{+1}_{00} \\ B'^{+1}_{11} \\ B'^{-1}_{11} \\ \cdots \\ B'^{\sigma}_{mn} \\ \cdots \end{pmatrix} \quad S = \begin{pmatrix} S\_HP_1 \\ S\_HP_2 \\ \cdots \\ S\_HP_L \end{pmatrix} \qquad \textbf{Équation 11}$$

La matrice C est appelée « **matrice de ré-encodage** ».

Le but du décodage est donc d'obtenir les signaux S_HP_i de façon à ce que le champ B' corresponde au champ ambisonique encodé B d'origine. Il faut alors veiller à ce que l'équation Équation 10 soit inversible.
On utilise alors une matrice dite « **matrice de décodage** » D telle que :

$$S = D.B \qquad \textbf{Équation 12}$$

La matrice D vaut alors :

$$D = pinv(C) = C^T.\,(C.C^T)^{-1} \qquad \textbf{Équation 13}$$

Dans le cas de système de reproduction uniforme, cette matrice de décodage est même simplifiée de sorte que l'on a :

$$D = \frac{1}{L} C^T \qquad \textbf{Équation 14}$$