Etude perceptive des inter-relations morphodynamiques gestes-images-sons

Arnaud Sicard

Stage de master ATIAM

effectué au LAM

du 13 Mars 2006 au 30 Juillet 2006

Encadrant : hugues Genevois

Remerciements chaleureux à Hugues Genevois et à toute l'équipe du LAM.

Table des matières

Table des figures

Introduction

Qui, aujourd'hui, peut affirmer qu'il n'est pas concerné par l'audiovisuel ?

Le cinéma et la télévision nous inondent à la fois d'images et de sons. Certains de ces ensembles sont naturels, c'est le cas pour des reportages télévisés : le son et l'image proviennent tous deux de la situation réelle. La restitution se trouve donc fidèle sauf en ce qui concerne les éventuelles techniques de prise de son et de capture d'image ou les détériorations des deux signaux. On s'attend à ce que les sources sonores et visuelles soient dans le précédent cas les plus adéquates, or si l'on vérifie la provenance des sources visuelles et sonores dans les films, elles sont pour la plupart distinctes, il suffit de l'exemple du coup de poing pris par Michel Chion dans son ouvrage l'Audio-Vision pour s'en persuader [CHION-90]. Avez-vous déjà « entendu » un coup de poing ? Il précise que sans le « bruit » ajouté, «on n'y croirait pas, les coups seraient-ils infligés pour de vrai ». Cet exemple peut se généraliser à beaucoup des phénomènes audiovisuels en partant du simple bruitage, jusqu'au doublage plus ou moins bien réalisé. Comment alors juger du meilleur ensemble, de celui qui va «marcher» le mieux ? La question se pose d'autant plus dans une situation de dessin animé ou de film d'animation dans lesquels chacune des parties doit être synthétisée ou empruntée à d'autres contextes.

Nous avons parlé jusqu'ici de cinématographie, les images sont le point de repère et on peut associer un son sans connotations à une image qui en comporte, mais le son peut lui aussi être le référent. On a donc de même pour un son donné, l'image qui peut changer, cela à condition que l'image ne soit pas signifiante. C'est par exemple le cas pour le VJing, c'est-à-dire la présentation d'images lors d'un concert, ou bien pour les PlugIns des lecteurs multimédias sur les ordinateurs. Un cas encore plus proche de nous est celui de la restitution de la parole, assistée ou non d'image de lèvres. Il est prouvé que la compréhension est nettement améliorée lors de la présentation de l'image. Ce type de relation audiovisuelle est utilisé dans des algorithmes de reconnaissance de parole, mettant en corrélation les données audio et les mouvements de la bouche et trouve des applications évidentes avec la visioconférence et les télécoms.

Tous les cas cités précédemment présentent au moins une des modalités ayant un contenu sémantique fort. Pour reprendre l'exemple du coup de poing, certains vont préférer un bruit de choc sec et pas trop fort et d'autres un son fort et gras. L'exemple pris juste avant permet de mettre en évidence deux choses : dans un premier temps le fait que ce soit un coup de poing nécessite un accompagnement sonore qui de toute évidence doit être percus-

sif mais aussi qu'il dépende de beaucoup de paramètres allant du type de film jusqu' aux préférences personnelles pour les fins connaisseurs en matière de coup de poing, et dans un second temps, les adjectifs utilisés pour décrire les préférences sonores sont très personnels et peuvent avoir des significations très différentes selon les individus. Les connaissances, les préférences ainsi que les habitudes de chacun jouent donc un rôle très important dans l'appréciation d'un ensemble. Cela ne fait qu'augmenter les difficultés de l'abord d'un tel sujet, pourtant nous sommes obligés de reconnaître que certains ensembles « collent » de manière universelle ; essayer de se dégager de ces contraintes pourrait permettre de comprendre pourquoi. L'unique moyen d'y parvenir est d'avoir recours à des objets audiovisuels ne signifiant rien tant du point de vue sonore que visuel pour le sujet, et c'est malheureusement strictement impossible puisque l'on cherche, par un procédé d'identification à rapprocher ce que l'on perçoit de ce que l'on connaît. On peut cependant essayer de limiter ces effets.

Pierre Schaeffer [SCHAE] propose dans son traité des objets musicaux de mettre entre parenthèses son savoir sur les sons afin de découvrir, dans sa perception, ce qui ne relève pas de l'interprétation ou de l'imagination. Il arrive aussi que l'on parle de vision réduite, il ne s'agit pas de voir des images en ayant les oreilles bouchées, mais de se forcer à ne rien décrypter de l'image qui atteit notre rétine, si ce n'est ses caractéristiques morphologiques. On adopte aussi volontiers ce type d'approche dans l'art pictural abstrait, il s'agit alors de percevoir les images comme ombres et couleurs animées, comme des "objets visuels" [LYON-98]. Peut-être pouvons-nous aussi créer le terme *« perception audiovisuelle réduite »*, en comparaison à *l'écoute réduite* et à la vision réduite, qui se concentrerait sur les qualités internes de l'objet audiovisuel. Dans certaines créations artistiques, il arrive que l'on rencontre des « relations audiovisuelles concrètes », les sons et les images créés forment un tout, et celui-ci n'a pas nécessairement, dans un tel cadre de création, de contenu sémantique.

Ces divers exemples montrent que définir le « meilleur » ensemble audiovisuel n'est pas simple ; d'une part, ce n'est pas nécessairement la situation réelle, quand elle existe, qui produit le meilleur résultat, et d'autre part, celui-ci dépend fortement, pour une image donnée, du son associé. De plus, la présence du contenu sémantique peut poser des divergences de points de vue, et même dans un cas qui en est dépourvu, l'auditeur cherche à identifier les sources visuelles ou sonores.

Sur ce sujet, la littérature est relativement restreinte surtout du point de vue scientifique. On trouve beaucoup de textes sur la question, mais ils restent très empiriques, quelques expériences ont été menées et peu de résultats ont été obtenus. La majeure partie de ces études concernent soit le cinéma, soit la parole, et malheureusement, les deux modalités perceptives sont rarement mises sur un même plan, en effet, dans le cinéma, l'image est un référent, et dans la parole, c'est le son qui est au centre. On se place donc dans une position où une des modalités est sensée « coller » à l'autre. L'emploi du terme « ensemble audiovisuel » me paraît biaisé puisqu'une modalité sert à appuyer l'autre. La création artistique permet de respecter une égalité des modalités au moins lorsqu'il s'agit de créer de l'image et du son au même moment, et par un même procédé. Comme un instrument de

musique produit des sons par des phénomènes physiques, il faudrait créer un instrument audiovisuel qui produise des ensembles audiovisuels par un mécanisme quelconque.

On peut répondre que de tels instruments existent déjà. Par exemple que les animations des lecteurs multimédias, par la lecture de son, la représente de manière visuelle par différents moyens, soit en affichant le spectre, le niveau sonore, etc...Je réponds que si une des modalités préexiste à l'autre, et qui faut extraire des paramètres au premier pour diriger le second, il s'agit plus d'une conversion que d'une synthèse bimodale. On peut aussi dire qu'un instrument de musique constitue déjà un instrument audiovisuel si l'on voit l'instrumentiste jouer. Je réponds cette fois que l'instrument de musique n'est que le mécanisme qui produit le son, qu'il ne constitue pas une forme en mouvement, et que donc le visuel n'existe ici pas dans le sens d'une partie d'un ensemble audiovisuel.

Lorsque l'on parle d'instrument, on sous-entend la présence d'un instrumentiste et une troisième modalité apparaît, la modalité haptique. Le geste vient donc compliquer encore un peu les choses, mais permet de définir plus précisément cet instrument audiovisuel et son mode de synthèse. On peut alors faire le parallèle direct entre un instrument de musique traditionnel et un instrument audiovisuel.

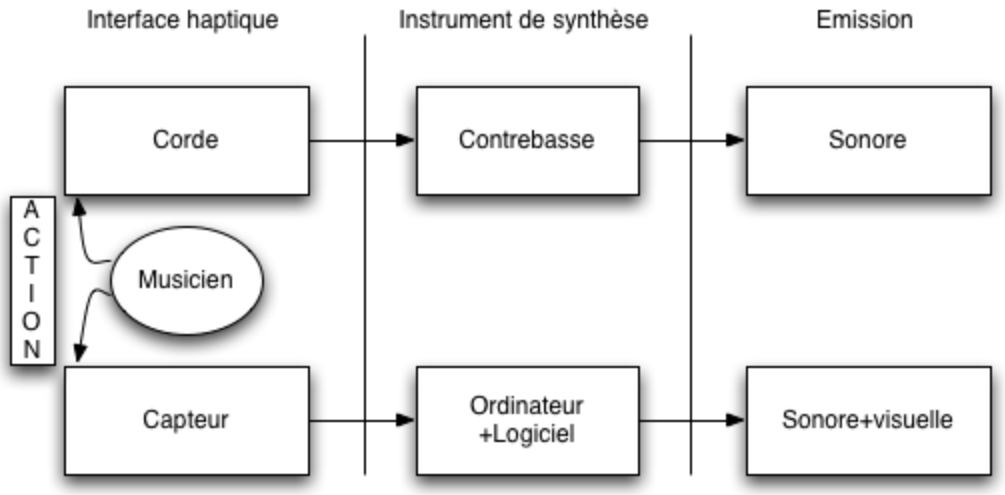

Fig. 1 – Instrument traditionnel-Instrument éléctronique

L'introduction du geste dans la problématique restreint davantage la bibliographie, et émettre des hypothèses devient nécessaire.

Nous nous sommes proposés de travailler dans un cadre de création artistique et de créer un tel instrument audiovisuel, les outils de création sonore et vidéo actuels étant aptes à fonctionner en temps réel. Cet instrument possède des réglages très variés permettant de donner diverses possibilités d'utilisation et ainsi de mettre en place une expérience de psychoacoustique. Cette plateforme très adaptable pourra aider à la vérification de lois déjà démontrées, mais cette fois dans des contextes différents, et éventuellement à mettre en évidence certaines lois nouvelles.

Je n'ai parlé jusqu'à présent que de « meilleur ensemble audiovisuel » et de « résultat » pour ne pas employer d'autres termes comme la fusion audiovisuelle, la cohérence ou bien la synchronisation des deux sources que je définirai plus précisément dans la suite.

Chapitre 1

Situation du projet dans le contexte général

1.1 Audiovisuel du point de vue physiologique

Selon Jacques PAILLARD [PAIL-92], il existe en physiologie un mécanisme qui cherche à extraire du flux d'informations reçues par les différents organes récepteurs celles qui sont cohérentes pour pouvoir définir un objet. Il y a donc séparation entre deux espaces perceptifs fonctionnellement distincts, à savoir un espace des lieux et un espace des formes. Ce principe est appliquable au visuel (le fond et la forme) et à l'audio (un accord et une mélodie par exemple) donc on doit pouvoir de même l'appliquer à l'audiovisuel.

Une équipe de chercheurs [BAUM-06] a étudié l'activité corticale liée à la perception visuelle logique de mouvement en présence d'une source de sonore, utilisant l'imagerie par résonance magnétique fonctionnelle (IRMf) (application de l'imagerie par résonance magnétique à l'étude du fonctionnement du cerveau. Elle consiste à alterner des périodes d'activité (par exemple bouger les doigts de la main droite) avec des périodes de repos, tout en acquérant des images de l'intégralité du cerveau toutes les 3 secondes.). Non seulement la taille des régions activées était sensiblement plus grande que pour le seul mouvement visuel ou auditif. mais plusieurs régions étaient activées uniquement en présence des deux stimuli simultanés.

Ces travaux justifient de manière expérimentale des idées avancées par la théorie de la Gestalt dont nous parlerons plus tard. Le fait que l'ensemble audiovisuel crée quelque chose de plus que la simple addition des deux stimuli permet d'exprimer que si il y a un lien qui est fait entre l'image et le son, alors les parties du cerveau concernées vont réagir. Cela pourrait être ne bonne manière de mesurer la qualité d'un ensemble audiovisuel.

1.2 Capteurs, mapping et lien gestuel

La différence entre les instruments traditionnels et électroniques est manifeste, ne serait-ce que par la notion de retour d'effort dynamique et bien que des travaux importants soient faits sur le sujet (Claude Cadoz et l'ACROE [ACROE]), les dispositifs l'intégrant sont assez rares. Cependant des efforts sont faits dans la direction des capteurs, souvent adaptés à un contrôle gestuel précis.

Serge de Laubier énonce dans un article [LAUB-06] le nombre de capteurs et leur variété utilisés sur la dernière version du Méta-Instrument (MI3), La position des avant bras se contrôle sans les mains, l'orientation de la poignée se contrôle par la paume, et chaque doigt comporte cinq capteurs indépendants. La mesure des rotations est au 1/20 de degré et la pression au 1/10 de gramme. , il fait mention de 54 variables continues indépendantes et simultanées, ainsi que de leur échantillonnage 16 bit transmis 500 fois par seconde pour l'ethernet et 100 fois pour le wifi pour des précisions pour les angles de $1/20^e$ de degré et de $1/10^e$ de gramme pour les pressions. Cette interface gestuelle est donc capable d'un contrôle très fin et extrêmement diversifié.

Un autre aspect sur lequel il met l'accent est le retour d'effort statique : les rotations ont une friction et une force de rappel réglables, et chaque touche est recouverte de mousse molle permettant d'apprécier la pression que l'on impose. Le retour d'effort dynamique serait à implémenter.

Le mapping est le seconde étape du fonctionnement d'un instrument électronique. Nicolas Montgermont [MONT] décrit dans son mémoire ces différents mappings. Il existe le mapping un a un, divergent, convergent ou bien une combinaison de ceux-ci. Les expériences de Hunt et al [HUNT] montrent que différents mapping créent des sensations très différentes pour l'instrumentiste. Il cite aussi Wessel [WESS] dont l'article met en évidence qu'un mapping simple propose une prise en main rapide de l'instrument et qu'un mapping complexe demande plus d'apprentissage, mais qu'à plus long terme l'instrument suscite plus d'intérêt. Ces considérations m'ont poussé à créer un mapping entièrement modulable pour l'instrument audioviuel.

Christian Jacquemin [JACQ-06] parle d'un mapping bidirectionnel entre image et son implémenté dans certains instruments multimédias, c'est a dire que certains paramètres du son agissent sur l'image et inversement. Il utilise de plus une construction de son instrument audiovisuel telle que le son et l'image sont tous les deux dirigés par les capteurs. Nicolas Mongermont dans sa synthèse Audiovisuelle avec PMPD (Physical Modellin For Pure Data) explique de même que beaucoup de systèmes de synthèse en temps réel s'appuient soit sur la vidéo pour créer le son, soit sur le son pour créer la vidéo. Ceci permet selon lui d'obtenir une cohésion forte entre l'image et le son.

Toutes ces considérations vont dans le sens de l'intégration de l'image à un place autre que celle qu'elle a tenu jusqu'à présent. Je cite Benoît courribet [COUR-05] « j'aspire à ce que la vidéo fasse partie de la musique, et non pas qu'elle en soit une illustration ou une représentation, au sens d'une trascription graphique ».

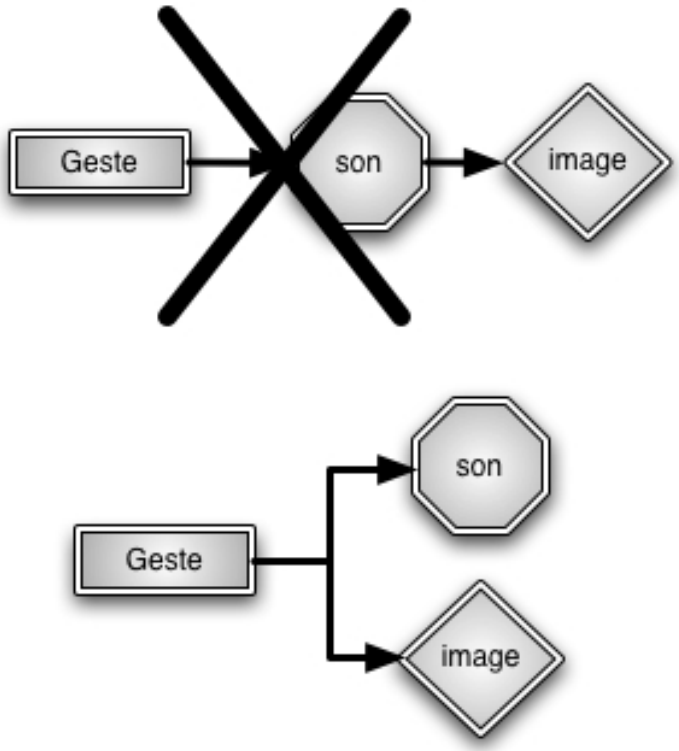

FIG. 1.1 – Mapping Audio/Vidéo

Il présente les outils de visualisation des lecteurs multimédia comme n'apportant rien au sonore, sinon une simple analyse, puisqu'ils ne font que le représenter différemment.

En accord avec ces discussions, l'instrument sera basé sur le schéma présenté plus haut.

1.3 Aspects spatiaux et temporels

1.3.1 Considérations empiriques

Comme je l'ai déjà souligné, les travaux de recherche comprenant une justification expérimentale de la qualité du lien Audiovisuel sont peu nombreux. Cependant, certains travaux comme ceux de Michel Chion [CHION-90] sont très complets et constituent une analyse assez fine du problème et la mise en évidence de lois importantes. Dans son ouvrage L'audio-vision, il présente une liste de concepts qui pourront nous être utiles dans notre travail.

Tout d'abord la valeur ajoutée, il s'agit de définir l'enrichissement de l'image par le son au point que l'information ajoutée paraît être présente dans l'image seule. C'est le cas par exemple dans la guerre des étoiles avec le son d'ouverture des portes de Ben Burtt tellement convaincant qu'il a suffit de filmer la porte ouverte puis fermée pour que l'auditeur la voie coulisser. Il précise que cet effet fonctionne surtout dans le cadre de la synchronie son/image, par un principe qu'il appelle synchrèse dont voici sa définition : *La synchrèse (mot que nous forgeons en combinant « synchronisme » et « synthèse ») est la soudure irrésistible et spontanée qui se produit entre un phénomène sonore et un phénomène visuel ponctuel lorsque ceux-ci tombent en même temps, cela indépendamment de toute logique rationnelle.*

Cette synchrèse permet la post-synchronisation, le bruitage et le doublage, et beaucoup de sources sonores différentes peuvent être associées à une seule image. Cet effet fonctionne aussi lorsque les images et le son n'ont rien à voir les uns avec les autres comme par exemple une syllabe prononcée et l'image d'un chien qui aboie. Il précise cependant que cette synchrèse n'est pas automatique car elle dépend du sens et s'organise selon des lois gestaltistes et des effets de contexte.

Cet effet peut aussi être influencé par les habitudes culturelles, mais semble avoir une base innée ceci étant prouvé par des réactions spécifiques de nouveaux nés à des stimulations synchronisées. De plus cet effet de synchrèse ne fonctionne pas par tout ou rien et est lié à une échelle.

Michel Chion semble donc donner une grande importance à la synchronisation temporelle, cet effet d'échelle pourrait donc être dû à des décalages de temps plus ou moins grand entre l'image et le son.

1.3.2 Expériences

Je me suis beaucoup inspiré de trois travaux qui utilisent des sources complémentaires et quelques fois redondantes, et qui méritent d'être décrites dans leur entièreté.

C. Nathanail dans sa thèse sur l'influence des informations visuelles sur la perception auditive [NATH-99] décrit des travaux concernant l'interaction spatiale entre image et son. Ces intéractions concernent les influences mutuelles sur la localisation des sources. THOMAS (1941)[THOM], et JACKSON (1953)[JACK] ont montré que la présence d'un stimulus visuel présent en même temps qu'un stimulus sonore fausse la perception spatiale de la source sonore et de plus que le type de stimulation visuelle influe beaucoup sur cette position estimée. Il apparaît aussi que lorsque la séparation angulaire augmente, l'association des deux stimuli est plus difficile et ainsi le biais diminue. De plus, l'influence de la localisation visuelle sur la localisation auditive est plus grande que l'influence réciproque et que la cohérence des stimuli augmente l'erreur du sujet. D'autres travaux traitent de cette influence d'une modalité sur l'autre, notamment sur l'orientation d'attention du sujet sur l'un ou l'autre des stimuli, de la possibilité qu'a le sujet à bouger la tête pendant l'expérience.

Il s'agit ici de localisation de source et non pas de qualité d'ensemble, et les résultats énoncés plus haut montrent que la provenance identique des stimuli tend à donner l'impression d'un même objet, il n'est pas nécessaire de plus détailler d'autres expériences.

Au niveau temporel, elle présente rapidement le fait que les humains perçoivent les décalages temporels avec une tolérance assez grande et qu'il existe un asymétrie dans les observations, en effet un retard du son sur l'image est bien mieux perçue que la situation inverse, ceci étant interprété comme l'habitude que nous avons de recevoir les stimuli sonores plus tard étant donné les vitesses de propagation inégales. Elle ajoute enfin que la nature des stimuli joue sur le caractère acceptable de l'ensemble, un bruit impulsif acceptera par exemple un plus grand retard que de la parole et que les réactions des sujets dépendent des consignes (Dixon et Spitz,1980) [DIXO].

Un travail bibliographique beaucoup plus complet est celui de Noël Château [CHAT-98] dans lequel il décrit les intéractions image/son à différents niveaux. Tout d'abord au niveau basique, Bertelson et Radeau (1981) [BETE] montrent un biais immédiat de a localisation de sources sonores et visuelles si elles ne coïncident pas. L'expérience la plus répandue pour mesurer ce biais consiste à présenter à un observateur une source brève de lumière (flashes de lumière) et une source sonore brève (pulsation sonore), et de demander à l'observateur de pointer la direction apparente de la source sonore dans le cas ou on lui demande de se focaliser sur la source lumineuse. L'expérience inverse peut être faite, et l'observateur doit se focaliser sur la source sonore (être en face). Voir figure 09876

Les résultats obtenus sont listés dans le tableau ci dessous.

Les résultats indiquent que la vision biaise l'audition plus que l'audition ne biaise la vision et qu'en pourcentage, ce biais diminue avec la séparation angulaire.

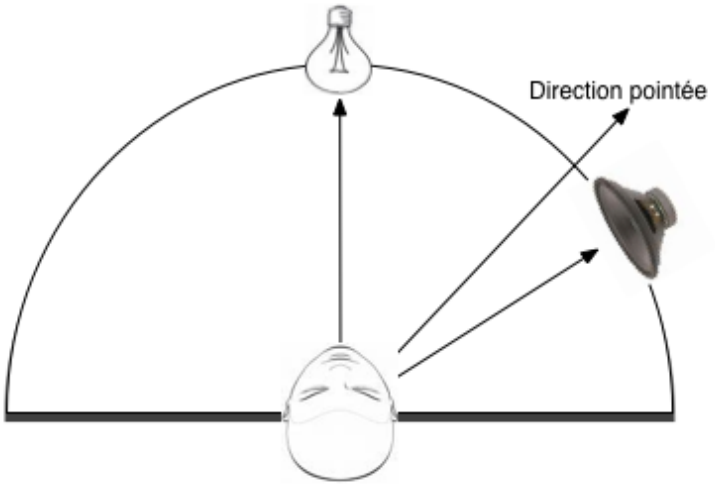

FIG. 1.2 – Expérience typique de décalages spatiaux

	V(A)		A(V)	
Séparation	Degrés	%	Degrés	%
7°	3.99°	57.0 %	0.34°	4.8 %
15°	6.27°	41.8 %	0.47°	3.1 %
25°	8.16°	32.6 %	0.61°	2.4 %

Tableau 1 : Biais immédiats absolus (en degrés) et relatifs (en pourcentages) de l'audition par la vision V(A) et de la vision par l'audition A(V) en fonction de la séparation angulaire des stimuli audiovisuels.

FIG. 1.3 – Biais Vision/Audition Audition/Vision

D'autres tests prenant comme base cet type d'expérience de séparation de sources ont été effectués. Notamment Thomas en 1941 [THOM] et ensuite Radeau et bertelson en 1987 [RADEA-87] qui y incluent le rythme. Ils ont montré que si les deux excitations utilisent le même pattern rythmique, le biais est plus grand que si ce n'est pas le cas. Si d'autre part le son est continu, des flashs lumineux suivant un pattern rythmique créent un biais de localisation, alors que le contraire est faux, ils justifient cela en disant que la vision est un sens fort et que l'audition est un sens faible. Radeau montre de plus que l'intensité des stimuli a elle aussi son influence, pour une même intensité acoustique, le stimulus lumineux le plus intense est aussi le plus biaisant, et d'autre part, pour une même intensité lumineuse, c'est le son le moins intense qui est le plus biaisé. L'influence d'autres paramètres a aussi été mesurée comme le temps de réverbération de la salle ou le rôle des facteurs conceptuels et attentionnels, mais nous ne nous y intéresserons pas dans le cadre de notre étude.

A un niveau non plus basique, mais cette fois en contexte semi réaliste, c'est a dire quand les stimuli ne son plus des flashes et des sinusoïdes, mais des images et des sons d'objets réels ou de personnes. Les résultats obtenus sont similaires à ceux obtenus dans un cas basique.

En ce qui concerne les interactions du point de vue temporel, elles n'ont pas été traitées à un niveau basique sauf dans un cas ou retards emporels et décalages spatiaux étaient mêlés, ce qui rend les résultats peu utilisables.

Daniel J. Levitin, Karon MacLean, Max Mathews and Lonny Chu dans leur article The Perception of Cross-Modal Simultaneity [LEVI] Expliquent que la plupart des travaux concernant l'asynchronie intermodale concerne l'audition et la vision. Les résultats présentent tous une asymétrie dans la précédence des modalités, la position synchrone se trouve majoritairement pour des retards de l'audio sur la vidéo. Ces mesures son donc toutes des mesures de seuil qui considèrent que l'on passe de non synchrone à synchrone a partir d'une certaine valeur de temps. Dixon et Spitz [DIXO] ont montré que les seuils étaient différents selon le stimulus, qu'il soit de la parole ou un exemple étranger au langage. L'expérience présentait donc des vidéos d'un coup de marteau, puis d'une personne parlant anglais. L'asynchronie se faisait ressentir pour des valeurs de retard du son de -75ms à +175ms, et pour la parole, de -130ms à +250ms. Cet effet serait lié à la façon de pronnocer qui présente le mouvement de lèvres avant que le son ne soit émis. Cette thèse est appuyée par McGrath et Summerfield [MCGR] qui ont fait passer la même expérience à des gens capables de lire sur les lèvres, les résultats donnent des valeurs de -65ms à +140ms, ce qui, je trouve, ne valide pas réellement l'hypothèse formulée, étant donné que l'amélioration peut être due à l'expertise des sujets. D'autres études ont obtenu des résultats pour des retards d'audio sur la vidéo de +90ms (Allan et Kristofferson [ALLA]), et de +100ms (Ganz [GANZ]).

Jaskowski [JASK] par un questionnement à choix multiple : avant, après, ou simultané avec un événement de référence, a obtenu un seuil de -65ms à +165ms. Les différents résultats sont présentés sur l'échelle de temps ci dessous.

Ces chercheurs ont donc cherché à mesurer différemment ces seuils en se plaçant en situation

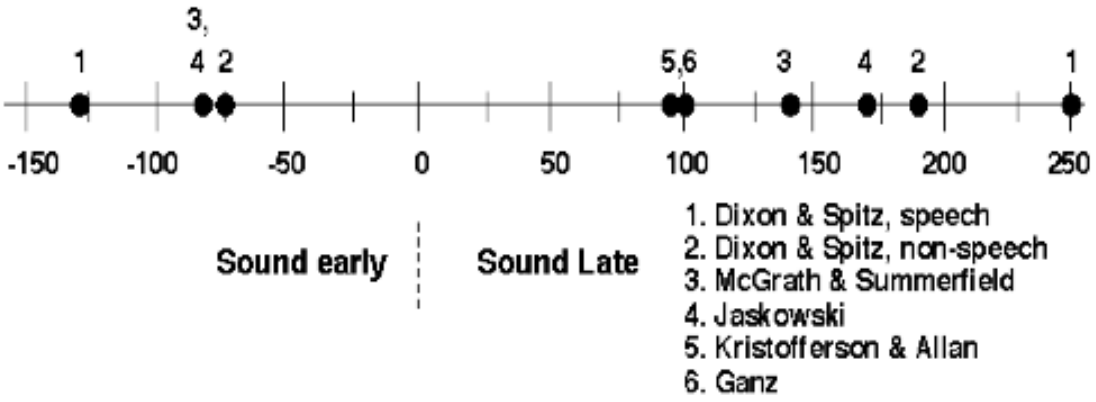

FIG. 1.4 – Récapitulatif des seuils de décalage temporel

réelle.

Les sujets étainent passaient l'expérience par paires. Un acteur et un observateur étaient désignés, et ils échangeaient leurs rôles tous les 90 essais. Ils disposaient en premier lieu d'un temps de pratique pour s'habituer à l'expérience, puis passaient 720 tests étalés sur trois jours. L'acteur frappe la surface d'un tambour avec une baguette, porte des écouteurs permettant une isolation phonique vis a vis du son réel et a les yeux bandés. On lui passe un son correspondant à celui du tambour mais avec un delai ; il ne reçoit donc que les informations de contact physique et son passé dans le casque. L'observateur est isolé dans une pièce située à deux mètres et séparée de la première par un double vitrage. Il reçoit le même signal auditif asynchrone que l'acteur lui aussi grâce à des écouteurs. L'observateur n'a donc accès qu'à la vue et son issu du casque. A chaque test, il est demandé aux sujets de dire si il y avait synchronie ou pas e d'ajouter un jugement (« pas sûr » « presque sûr » et « sûr »).

FIG. 1.5 – Schéma de l'expérience de la baguette

Les résultats indiquent que si l'on considère valables les valeurs citées comme asynchrones 75% du temps, on obtient des temps de -25ms à 42ms pour l'acteur et -41 à +45 pour l'observateur. Ces temps sont bien inférieurs à ceux obtenus dans les autres expériences.

On constate de plus qu'un auditeur est moins sensible à la simultanéité qu'un acteur, cela semble bien s'accorder avec le problème de la latence pour les instruments de synthèse.

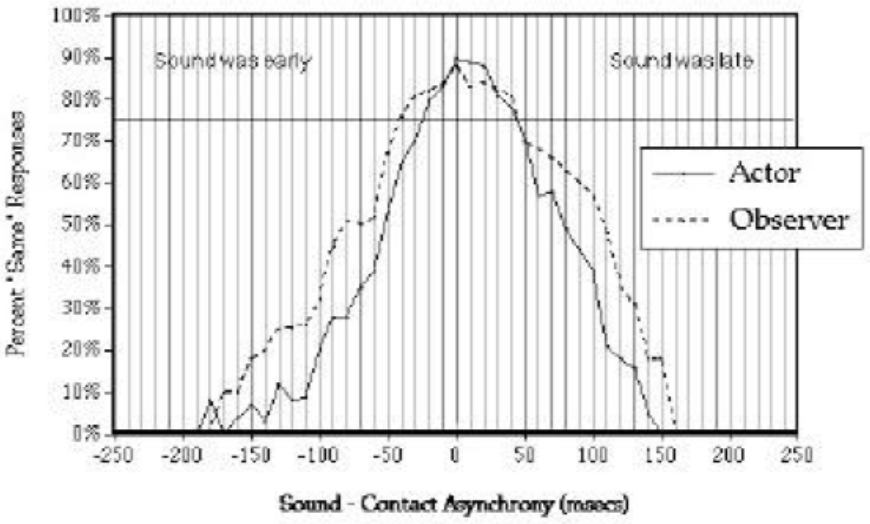

FIG. 1.6 – Résultats de l'expérience de la baguette

Pour cette équipe de chercheurs, la simultanéité est fortement liée à l'anticipation. Celle-ci est ici renforcée par le geste de l'acteur. En situation devant un écran, même l'expérience avec le marteau n'arrive pas aux mêmes résultats. Le réalisme de la situation audiovisuelle joue donc en la faveur d'une synchronie plus précise, et en ce qui concerne la synchronie geste-son, elle paraît encore plus resserrée. Ils se sont ici arrêtés à des expériences bimodales, mais auraient aussi pu essayer du trimodal en enlevant simplement le bandage sur les yeux de l'acteur.

Van de Par, Steven,Kohlrausch, Juola et James [VAND] ont cherché à mettre en évidence le rôle du type de mesure sur la détection de l'asynchronie. Il existe en effet trois méthodes pour la mesurer : la première consiste à donner trois catégories, le son avant, synchrones, et l'image avant, la seconde ne demande que si c'est synchrone ou non, et enfin la troisième demande quelle modalité, audio ou vidéo précède l'autre. Il s'avère que les deux premières méthodes semblent équivalentes et relativement robustes, donnant en moyenne une synchronie perceptive de 35ms, mais que la troisième induise des stratégies de décision qui diffèrent selon les sujets. Les points de synchronie observés pour ce type de réponses donnent pour certains sujets des valeurs de synchronie perceptive négatives, c'est a dire que l'audio est en avance sur la vidéo. Ces résultats entrent en contradiction avec toutes les expériences effectuées sur le sujet, remettant en cause la validité des conditions d'expérience et insistant sur l'importance du mode opératoire.

Jusqu'ici, le principe des expériences rencontrées est de présenter deux sources et d'expliquer au sujet qu 'elles sont distinctes, on cherche alors l'influence de l'une sur l'autre. Il me semble plus intéressant, ou du moins ce qui se rapproche de ce que l'on cherche, de partir

du principe que l'on a une source audiovisuelle qui fonctionne, c'est a dire que bien que les deux stimuli ne soient pas issus de la même source, on a l'impression d'un seul ensemble, et, en introduisant des distorsions, d'essayer de remarquer les dégradations perceptives associées. Ce genre d'approche a été utilisée pour l'étude de l'effet du ventriloque, c'est a dire lorsqu'une personne réussit à parler sans bouger la bouche et que la voix semble appartenir à la marionnette qu'il remue en même temps. N. Château explique à ce sujet que le terme d'effet du ventriloque est souvent associé, à tort, à tous les cas dans lesquels une source sonore est attirée par une source visuelle. Or, selon lui, et je partage son opinion, on ne peut utiliser ce qualificatif que si, en tant qu'observateur, nous ressentons une « fusion perceptuelle entre son et image ».

On peut aussi dans ce cas faire une distinction entre les intéractions spatiales et temporelles.

Thurlow et jack en 1973 [THRU-73] ont testé l'influence sur l'effet du ventriloque de la séparation angulaire dans un plan horizontal. L'image d'une personne parlant était présentée sur un écran et le son diffusé sur une enceinte décalée d'un angle alfa et ceci en environnement anéchoique. Les conclusions sont que l'effet du ventriloque reste fort jusqu'à des angles de 30 degrés, mais qu'il chute nettement à 40 degrés jusqu'à devenir inexistant au delà de 50. Ils ont de plus testé le même effet mais cette fois en environnement réverbérant, la décroissance de l'effet se trouve réduite et 10 degrés supplémentaires sont à ajouter à chaque limité présentée plus haut. Dans un plan vertical, l'effet est encore plus robuste face à la séparation angulaire et des décalages de 55 degrés en environnement anéchoique, et de 195 degrés en environnement réverbérant.

Les intéractions temporelles sont aussi étudiées par (Jack et Thurlow 1973) toujours avec le même dispositif expérimental. Leurs résultats sont que 100ms de retard du son sur l'image sont parfaitement acceptés, que 200ms déteriore considérablement l'effet et que 300ms le réduit à néant. Cavé en 1992 [CAVE] a aussi mesuré cet effet mais cette fois dans un contexte de salle de cinéma à 7m d'un écran (les 20ms de retard dus à cette distance sont pris en compte dans les résultats) des observateurs devaient dire si face à la projection d'un clap de cinéma et du son correspondant, le son venait « en avance (1)», « en retard (2)», ou « au bon moment (3)». Vingt quarte observateurs jugeaient chacun 10 retards. Les résultats obtenus donnent la courbe suivante :

On peut remarquer sur cette figure que pour la valeur 2, qui représente la perception de simultanéité des signaux, le retard réel du son n'est pas nul, mais aux environs de 40ms. Il existe de plus une zone dans laquelle la pente est assez faible (entre +20 et +100ms) qui est perçue comme synchrone. De plus la pente est moins forte pour des valeurs de retard du son que pour des retards de l'image, ce qui conforte l'hypotèse de l'asymétrie concernant notre préférence à des retards de son plutôt que d'image.

Il est quelque fois fait mention des lois de la théorie de la Gestalt au sujet des rapports image/son, mais jamais avec des explications précises. Noel château explique que la dynamique des évolutions temporelles en vision est régie par les lois de la Gestalt et en particulier la loi de destin commun (ou *common fate*) c'est d'ailleurs aussi le cas pour l'audition. Cette loi explique que des éléments évoluant de la même manière au cours du temps

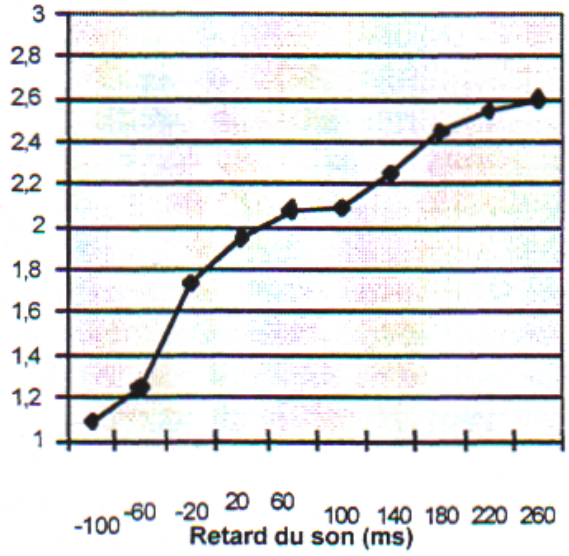

FIG. 1.7 – Décalages temporels en condition cinématographique

ont très peu de chances de provenir de sources différentes et sont alors groupées en un flux unique. Peut être peut-on étendre cette loi à des ensembles audiovisuels. Radeau ainsi que O'Leary et Rhodes en 1984 [OLEA] ont formulé cette hypothèse. Jack et Thurlow ont mis au point une expérience visant à la vérifier et de plus à montrer l'influence du degré de réalisme d'une expérience. Une marionnette à main simple a laquelle on pouvait rajouter des éléments (nez, yeux, etc...) et dont la mâchoire était mise en mouvement était placée face à un auditeur. Le son était émis par un haut parleur caché et formant un angle de 20 degrés avec la marionnette.

Voici quelques remarques succŽdant la lecture des articles :

- Les résultats montrent que le degré de réalisme et la cohérence des paroles avec les mouvements de la marionnette ont leur importance, mais que ces derniers étaient beaucoup plus influents.
- La justification de l'hypothèse des lois de la Gestalt par cette marionnette ne me paraît pas très convaincante.
- A la lecture de ces articles, il semble nécessaire que certains éléments soient présents dans l'instrument audiovisuel ainsi que dans les expériences.
- La diversité des capteurs peut permettre plusieurs contrôles gestuels et augmenter ainsi l'expressivité.
- Le cadre audiovisuel oblige a suivre un schéma de mapping dans lequel audio et vidéo sont contrôlés directement par les mêmes capteurs.
- Le besoin de réalisme nécessite une bonne représentation, et on doit se dégager au maximum de la sémantique.
- Le point clé du problème semble être lié aux aspects temporels, et puisque cela ne sera pas l'objet de notre étude, la cohérence spatiale doit être respectée.

1.3.3 Quelques définitions

Certains termes sont employés dans les divers articles que j'ai pu lire, et il me semble indispensable de les définir plus précisément afin de les utiliser à juste titre. Les définitions sont extraites de Wikipédia [WIKI] et du Dictionnaire de l'Académie française en version informatisée [DICAC].

SYNCHRONISME : *Le synchronisme désigne le caractère de ce qui se passe en même temps, à la même vitesse. L'adjectif synchrone définit deux processus qui se déroulent de manière synchronisée.*

On peut constater que le synchronisme bien que tout le monde comprenne de quoi il s'agit, il n'est utilisé que dans la moitié de sa définition, la notion de vitesse n'est jamais utilisée. J'utiliserai donc dans la suite le terme de la même façon, c'est a dire sans notion de vitesse.

COHÉRENCE : *Emprunté du latin cohaerentia, « connexion, adhésion ».*

Union étroite entre les éléments constitutifs d'un corps. Cohérence spatiale, temporelle. Synonymes : Liaison étroite, adhérence mutuelle, agrégation, connexion, harmonie, rapport logique.

Cette notion introduit donc la coexistence de deux entités très fortement liées, mais restant séparées pour autant, il n'y a pas deux éléments qui finissent par n'en former qu'un.

FUSION : *D'une manière générale, le mot fusion (du Latin fusio, du verbe fundere qui signifie fondre) désigne l'action consistant à faire d'une ou plusieurs entités une unique entité. Plus particulèrement, le mot est employé dans plusieurs domaines. 1. BIOL. Phénomène par lequel deux ou plusieurs éléments s'associent en une structure unique. La fécondation provient de la fusion de deux cellules appelées gamètes. 2. PHYS. NUCL. Formation d'un noyau atomique à partir de deux noyaux plus légers. La fusion nucléaire dégage une très grande quantité d'énergie. 3. Réunion en un ensemble de deux ou plusieurs éléments. La fusion de deux territoires, de deux provinces. La fusion de deux partis politiques. DROIT COMMERCIAL. Opération par laquelle deux ou plusieurs sociétés mettent leurs biens en commun pour n'en plus constituer qu'une seule. Opérer la fusion de deux entreprises. Procédure de fusion. Contrat de fusion. Fig. Alliance intime. La fusion des cœurs, des esprits.*

Le terme de fusion peut donc s'appliquer à l'audiovisuel lorsque les parties audio et visuelles seules n'existent plus, seul l'objet audiovisuel reste présent.

Chapitre 2

Mise au point de l'expérience

2.1 La simplification

Lors de la présentation du stage, les idées d'études étaient relativement floues et un des points de départs était un patch MaxMSP/Jitter présentant une voile en trois dimensions implémentée à l'aide de la librairie PMPD (Physical Modelling for Pure Data de Cyrille Henry [HENR-04] et portée sous mac os X par Ali Momeni).

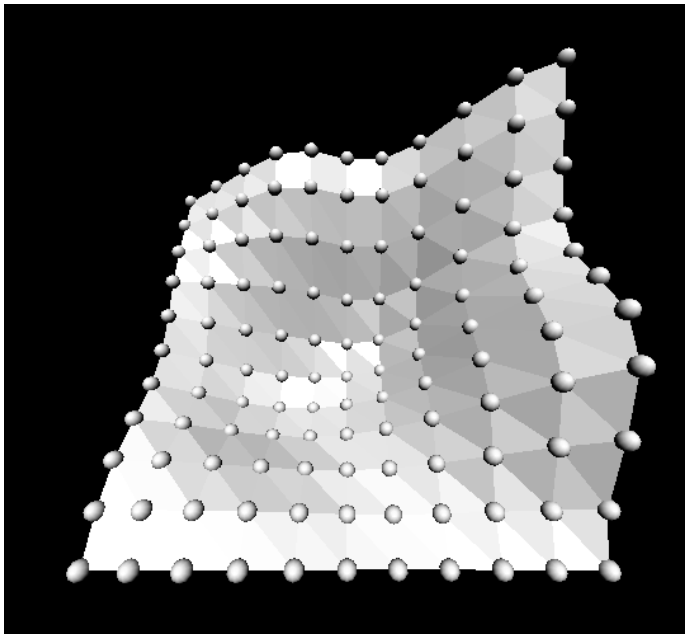

FIG. 2.1 – Voile réalisée avec PMPD

Cette voile pouvait connaître des déformations en temps réel, et il était donc simple de diriger son comportement à l'aide d'un capteur.

Deux problèmes m'ont pourtant poussé à abandonner cet outil : cet objet est très complexe, il comporte des textures, et une simple excitation ponctuelle entraîne un mouvement ample et étalé dans le temps (qui peut être amoindri en augmentant l'élasticité, mais pas suffisamment), identique à la réaction qu'aurait une voile sur laquelle on mettrait un coup sec. A cela s'ajoute l'aspect sémantique, en effet tout le monde à son idée du mouvement que doit avoir une voile, et aux vues de la bibliographie, il était nécessaire de l'éviter le plus possible.

Suite à une présentation au LAM par Cyrille Henry de cette librairie de synhèse par des modèles physiques, j'ai pu me rendre compte qu'il l'utilisait du point de vue graphique comme d'une représentation de l'objet physique rayonnant, comme un moyen d'étendre l'instrument électronique à autre chose que le contrôleur, l'ordinateur et le logiciel, lui ajoutant aussi une forme propre, graphique.

Nous avons discuté de la place que nous voulons donner à l'image dans une relation audiovisuelle. L'instrument de musique possède certes un caractère graphique, mais son mouvement visuel n'est pas à lier au sonore. Pour citer un exemple, le trémolo du violon ne se voit pas sur le violon en lui-même, les changements vibratoires des modes de la caisse de la caisse sont imperceptibles, or du point de vue sonore, la variation est largement audible. Les créations de Cyrille Henry sont cependant très intéressantes autant du point de vue scientifique qu'artistiques, mais l'utilisation de la librairie PMPD semble donc inadaptée pour l'étude que l'on veut mener au regard de l'idée audiovisuelle que nous avons décidé de suivre.

Il nous fallait donc créer un outil de synthèse à la fois graphique et sonore, évitant au maximum les relations d'ordre sémantique, pouvant être dirigé par des capteurs gestuels et suffisamment modelable pour faire les tests sur de décalage temporel.

2.2 L'instrument audiovisuel

2.2.1 Les capteurs

Comme nous l'avons spécifié plus haut, nous avons cherché à diversifier les types de capteurs et ceux que nous avons utilisé sont les suivants :

– Un capteur de pression (FSR, Force Resistive Sensor) : Ces capteurs modifient la résistance selon la pression appliquée. La pression d'un doigt sur le capteur de 10 g à 10 kg entraîne la baisse linéaire de la résistance de 2 MOhms à 1 kOhms. Ils sont de plus très fins (0,3 mm).

Ce capteur nous permet donc d'avoir un réponse assez sensible pour un petit geste. Nous avons ajouté de chaque côté du capteur des coussinets de feutre pour produire un retour d'effet statique et un confort d'utilisation (appuyer sur une surface due finit par faire mal au doigt...).

FIG. 2.2 – Capteur FSR

– Un capteur de distance. Dans un premier temps, nous avions choisi un capteur fonctionnant sur la base de la triangulation optique en technique IR sharp dont voici une image :

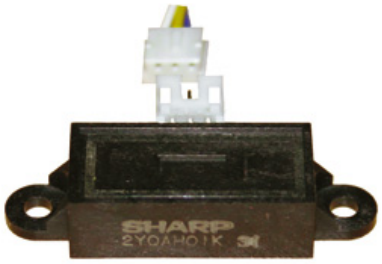

FIG. 2.3 – Capteur de distance infrarouge

Ce capteur délivre une tension en fonction de la distance au premier objet que le faisceau infrarouge recontre. Cela posait le problème du matériau de la surface en question et de plus, la tension délivrée n'est pas monotone en fonction de la distance. Nous avons donc préféré utiliser un autre type de capteur, une photorésistance :

Le geste pour ce type de capteur est beaucoup plus ample, pour de bonnes conditions de luminosité (lumière diffuse), les extrema sont atteints pour des valeurs allant de 0 à 1m.

– Pour faire le lien avec l'ordinateur, nous avons utilisé un capteur transformant des tensions et signaux midi, le miditron :

Cette interface permet d'alimenter 10 capteurs du type de ceux que nous avons décrit, mais aussi de transformer des signaux midi en tensions et de les envoyer à dix autres capteurs.

On peut l'alimenter avec une pile 9v ou par le secteur.

Un patch max permet de programmer les canaux midi utilisés et leur liaison avec les entrées/sorties ainsi que de régler les extrema de tension correspondant aux valeurs midi

FIG. 2.4 – Photorésistances

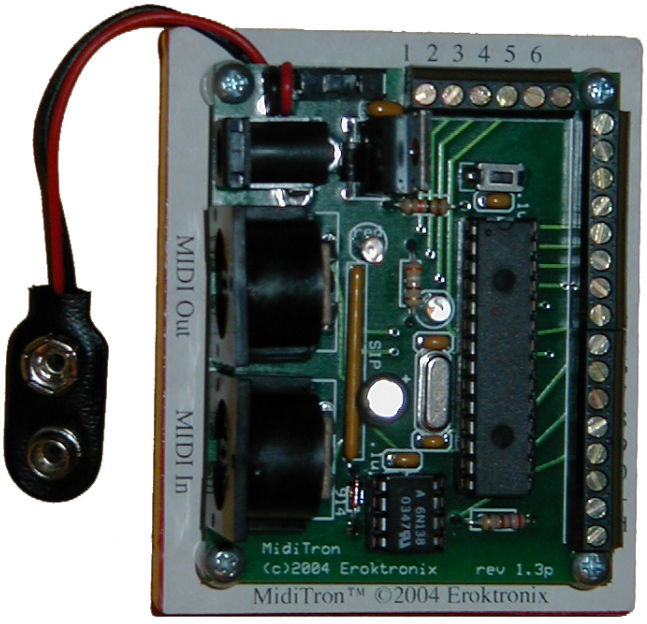

FIG. 2.5 – Miditron

(0V→0 et 5V→127 ou 1V→0 et 3V→127). Le lien midi avec l'ordinateur était fait à l'aide d'un convertisseur midi vers usb.

– Nous avons aussi utilisé une tablette graphique wacom ajoutant un capteur à dimension spatiale et pour laquelle il existe un external max permettant de recevoir les valeurs de position et de pression délivrées par la tablette écrit par Jean-Michel Couturier du LMA (http ://www.jmcouturier.com/download.html).

Voici des photos de mon « bricolage » final :

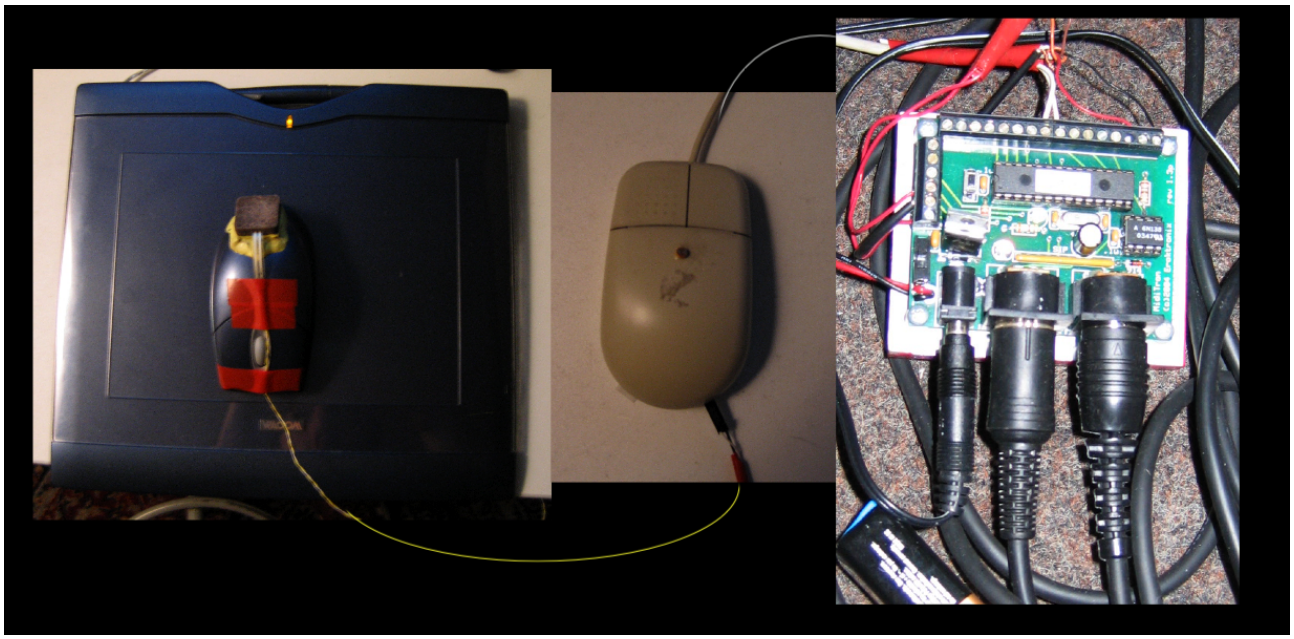

FIG. 2.6 – Ensemble des capteurs et du montage

A gauche, le capteur de pression attaché avec de la Patafix à la souris de la tablette graphique, il est de plus branché à une souris recyclée (au milieu), démontée et servant de réceptacle aux circuits électriques (2 ponts divisieurs de tension pour passer des variations de résistance aux variations de tension) et à la photorésistance. Le tout est ensuite relié au miditron puis à l'ordinateur.

–

2.2.2 Max MSP/Jitter

Max/MSP (Cycling74) ainsi que son équivalent libre Pure Data sont deux logiciels conçus par Miller Puckette dans les années 80. Ce sont des logiciels de programmation graphique de traitement du signal audionumérique. L'environnement de travail s'appelle un patch (c'est une feuille blanche), sur lequel on dépose des boites qui représentent des fonctions, puis ces boites sont reliées entre elles selon les entrées et sorties qu'elles comportent. Max

MSP est un logiciel dédié à la création sonore et musicale, il intègre la norme midi, et est associé à Jitter pour la synthèse graphique qui fonctionne sur le même principe.

Voici donc le patch que j'ai écrit pour réaliser cet instrument, il est divisé en plusieurs parties, les différentes synthèses sonores, la synthèse graphique, le réglages des retards, l'interface de mapping modulable, la réception des valeur des capteurs ainsi que leur enregistrement. Je vais présenter quelques parties de celui-ci ainsi qu'un schéma général.

On peut remarquer que ce sont les mêmes valeurs qui contrôlent l'image et le son et que la synthèse audio est indépendante de la synthèse vidéo.

Ci dessous, le patch de la synthèse FM,

On contrôle ici l'intensité totale (intens2), la hauteur de la sinusoïde (hauteur2), la hauteur de la porteuse (brillance2) ainsi que la valeur de l'indexe de modulation (niveaubrillance2). Quatre instruments ont été créés :

- Une sinusoïde de relativement basse fréquence plus ou moins enrichie de manière harmonique par synthèse additive et dont les harmoniques varient en [f061]/n, les paramètres variables sont l'intensité globale (cele ci possède un déclenchement fonction de la dérivée de la valeur reçue, si l'évolution est trop lente, le son n'apparaît pas), la fréquence fondamentale, l'enrichissement spectral (fonction de [f061]) et la spatialisation,
- Une lecture de quatre samples liés à la spatialisation et sur lesquels étaient appliqués deux filtres, un passe bas et un passe haut de même fréquence de coupure, les paramètres étaient donc le niveau sonore, la fréquence de coupure des filtres, leur coefficient de qualité, la balance entre ces deux filtres, et la spatialisation donc par la même occasion la sélection du sample (les samples sont disosés comme suit dans un plan xy carré limité : sample de gouttes d'eau(-1,-1), sample de vent(1,-1), sample de rivière(1,1), sample de feu(-1,1)), les samples ont été téléchargés sur le site freesound (http ://freesound.iua.upf.edu/).
- La synthèse FM que l'on a vue précédemment,
- Une synthèse additive de bruits blancs filtrés et dont les harmoniques sont impairs et décroissants en alfa/n, les paramètres étaient cette fois le niveau sonore, la hauteur de la fondamentale, et l'enrichissement spectral (fonction de alfa).

Tous ces paramètres sont reçus dans ces patchs après avoir été routés au niveau du panneau de contrôle dont voici un aperçu du mapping modulable.

On peut reconnaître les quatre paramètres cités dans le patch de synthèse FM, plus deux autres correspondant à la spatialisation.

Ainsi chaque paramètre de synthèse peut recevoir les données d'un des capteurs, et ajuster à la main la courbe de son évolution.

Pour la partie graphique, j'ai décidé d'utiliser une forme simple, la sphère, et d'utiliser au maximum les changements que l'on peut lui faire subir dans un cadre de rendu en 3D.

J'ai choisi la position dans l'espace en x, y et z (un plan référent vu en perspective permet de séparer la position dans le plan vertical et la hauteur de la sphère), la couleur, la taille, ainsi que le nombre de dimensions la définissant (dans Jitter le rendu 3D d'une sphère

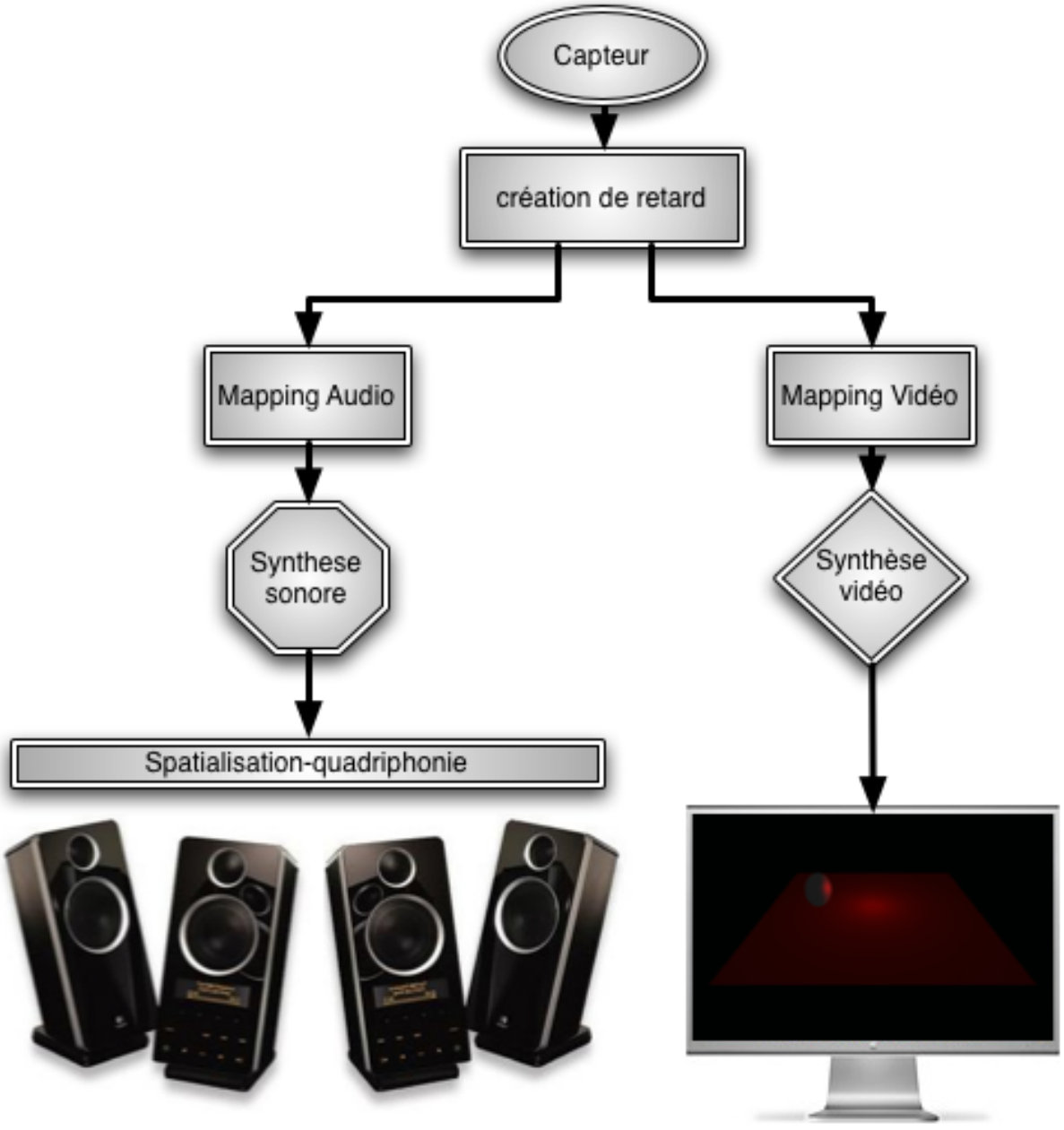

FIG. 2.7 – Schéma du patch MAX

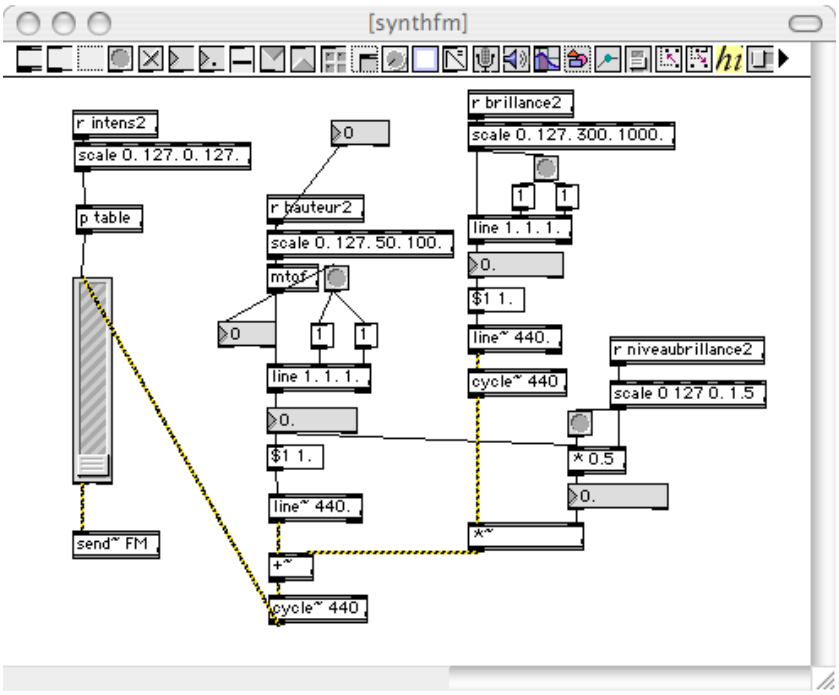

FIG. 2.8 – Patch de synthèse FM

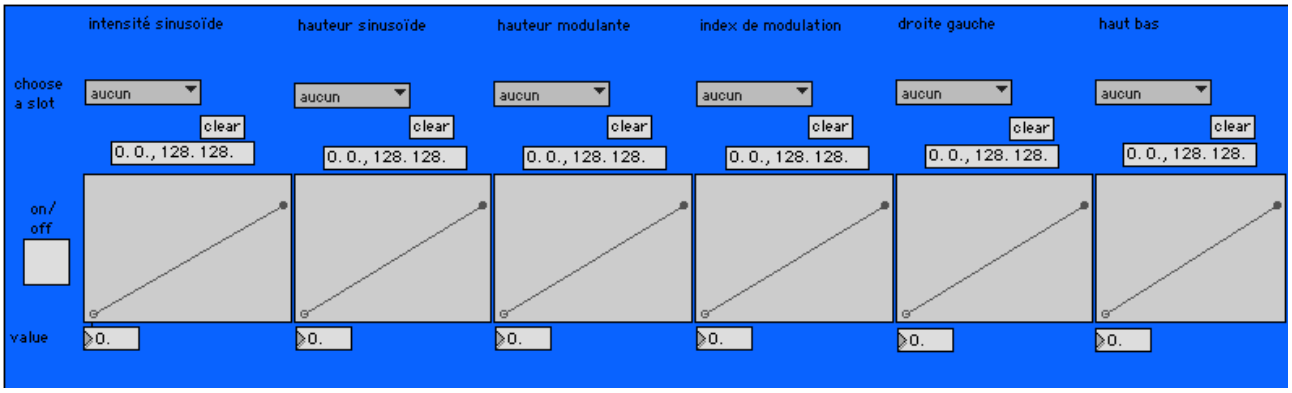

FIG. 2.9 – Partie du patch de mapping de la synthèse FM

dépend de ce nombre, une sphère à trois sommets correspond à un triangle, à quatre à une pyramide, et ainsi de suite jusqu'à former une vraie sphère). Voici la partie du patch concernant certaines des modifications de la boule :

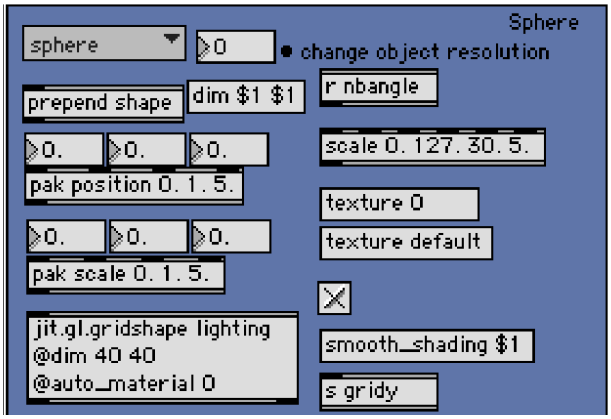

FIG. 2.10 – Patch de création graphique

Le résultat graphique se présente comme ci dessous :

Le plan en perspective et les effets d'ombres servent à améliorer le côté réaliste. D'autre part la lumière vient du centre de ce plan et sert à se représenter soi même dans la situation, au centre de ce plan. Au départ, les axes x et y étaient liés à la spatialisation (quadriphonie) ce qui explique d'autant mieux le choix du plan horizontal.

Pour l'utilisation de l'objet et les expériences, nous avons utilisé un powermac G5 2x2 ghz donc amplement suffisant en ce qui concerne les calculs (disposé à l'extérieur de la chambre d'écoute pour minimiser les nuisances sonores, et les rallonges pour ce type d'écran sont très chères. . .). Pour l'audio, nous avons utilisé une carte son firewire Mobile IO nous permettant de faire le lien avec la chambre d'écoute du LAM dans laquelle sont installées 12 enceintes adaptées à la restitution d'environnements sonore. Nous n'avons utilisé que quatre d'entre elles. Un micro était disposé dans la chambre et un autre à l'exterieur pour permettre un contact vocal pendant les expériences. Pour la restitution visuelle, nous avions besoin de deux écrans, un pour le sujet, et un à l'extérieur de la chambre, nous disposions d'un Cinéma HD display 23 pouces d'apple pour le sujet et d'un autre de taille plus raisonnable pour contrôler la manipulation.

Voici une image de l'ensemble du montage en situation :

2.2.3 Problèmes de mise en œuvre

Un souci important que j'ai rencontré assez tôt a été la manière d'éclairer la photorésistance. En approchant la main au dessus de celle ci, on peut faire varier son éclairement. Seulement si les sources de lumière sont multiples ou de nature non diffuse (une ampoule simple est non diffuse lorsque l'on se place à 1,50m, l'ombre d'un objet à une telle distance possède

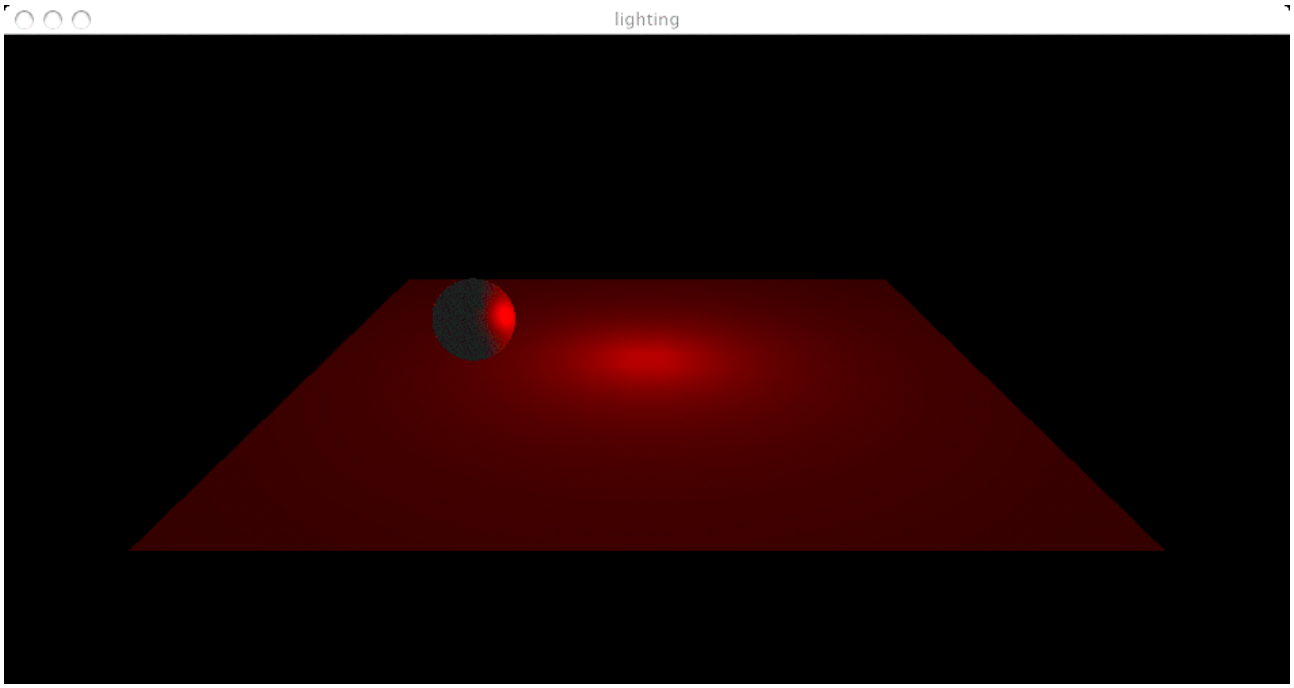

FIG. 2.11 – Capture d'écran du rendu graphique

des contours très nets !), la loi d'évolution n'est pas uniforme, mais elle ne l'est pas non plus lorsque la lumière est diffuse. L'écran lui-même fournissait une source de lumière, ce que j'ai réduit en changeant la couleur du fond (elle était blanche au départ). La solution a donc été de diriger une lampe contre un des panneaux en tissus, ce qui évitait l'éclairement direct et ne créait pas non plus une lumière trop diffuse. En choisissant bien la direction selon laquelle on lève la main au-dessus du capteur, on arrive à obtenir une évolution quasiment linéaire.

Un autre problème de taille s'est posé, le powermac que nous avons utilisé n'est équipé que d'une carte graphique de qualité moyenne (NVidia GeForce FX 5200 64MB) et pas du tout adaptée à du double écran dont un de 23 pouces utilisant des applications 3D. L'optimisation du patch a donc été de rigueur, et quelques artéfacts mineurs (inexistants sur un simple écran) ont été inévitables. Les nombres de couleurs utilisés par les écrans ont donc été diminués et la résolution sur l'écran de contrôle en a pâti elle aussi.

Le passage par la norme midi grâce au miditron permet d'une part de faire le lien entre de l'électronique et l'ordinateur, mais il réduit d'autre part le flux de données qui passe de continu à un débit midi, et transforme des valeurs continues en valeurs entières de 0 à 127 ce qui n'est pas sans conséquences sur la fluidité des images. De plus, le miditron, en tous cas avec les capteurs dont nous disposions créait des interférences entre les deux capteurs. Par exemple si le captur de pression était laissé au repos, et que de la lumière était émise sur la photorésistance, la valeur reçue par max variait constament et rapidement entre 0 et 2 posant un problème de réception permanente de données par Max.

FIG. 2.12 – Photo du dispositif expérimental

Nous aurions pu résoudre ce problème en utilisant un nouveau système usb http ://www.create.ucsb.edu/ dano/CUI/ ayant exactement les mêmes caractéristiques que le miditron, à ceci près qu'il est directement alimenté via usb, qu'il envoie des valeurs continues, et qu'il coûte quatre fois moins cher... Une version wifi existe d'ailleurs depuis peu permettant la mobilité. Je n'ai pris connaissance de cet objet qu'une fois l'expérience déjà lancée.

J'ai dû de plus transférer le patch sous Macintosh, l'ayant commencé sous Windows, ce qui représente l'entière révision et un grand nombre de changements, comme par exemple tous les liens avec les capteurs qui sont faussés, les patchs associés à la tablette graphique et au miditron étant spécifiques au système d'exploitation.

La réalisation de cet outil m'a pris beaucoup de temps, je ne connaissais auparavant pas le logiciel,. Les patchs d'aide, la documentation et les forums de MaxMSP/Jitter m'ont donc été d'une aide précieuse.

L'expérience pilote

2.3 L'expérience pilote

Les outils étaient enfin tous prêts, et j'ai décidé de procéder à une expérience.

Ma première idée était la suivante : je voulais, suivant un procédé adaptatif, trouver chez des sujets la valeur de décalage temporel entre son et image qui leur paraissait correspondre à la simultanéité en situation d'ovservateur puis d'acteur. J'ai manipulé un peu les capteurs dans une configuration précise du mapping tout en enregistrant les flux de valeurs. Max était donc capable de refaire la même chose sans que je sois présent. L'ajustement était possible avec les flèches du clavier et a chaque fois que le sujet avait l'impression de synchronie, la valeur était stockée et un autre déroulement audiovisuel était présenté, la valeur de départ de décalage étant aléatoire.

Je voulais ensuite mesurer l'influence de la synchronisation dans la réalisation d'une tâche. Le sujet devait à l'aide des capteurs et utilisant la sinusoïde reproduire « au clair de la lune », les décalages étaient aléatoire a chaque itération et je relevais le temps d'exécution et enregistrait la séquence pour mesurer la justesse.

Enfin j'avais projeté d'associer toutes les évolutions des images à toutes celles des sons et d'essayer de catégoriser ces ensembles à partir de jugements de valeur et de verbalisation du sujet.

Lors du passage du premier sujet, j'ai été confronté à une remise en question importante. Malgré les retards du son ou de l'image atteignant parfois 300ms, le sujet était tout à fait convaincu de la bonne fusion audiovisuelle, et il a été surprenant de voir que celui-ci partant par exemple d'un retard sonore de 200ms continuait à augmenter ce décalage, pour finir en me disant qu'il était incapable de quoi que ce soit. Il a ajouté que quelques fois, il percevait un décalage lors d'évènements percussifs, mais était incapable de régler le système. Je n'ose

imaginer la réaction d'incompréhension qu'aurait éprouvé un sujet auquel je n'aurais pas spécifié qu'il s'agissait de décalages temporels.

Suite à ce premier échec, j'ai souhaité continuer l'expérience en proposant un second mapping, mais le résultat a été le même. Les deux autres tests n'avaient alors plus vraiment de sens, et j'ai décidé d'abandonner ce type de test. S'il fallait présenter uniquement des stimulations percussives pour obtenir un résultat, alors autant se référer directement à la bibliographie.

Chapitre 3

Expérience définitive et résultats

3.1 Idées préalables

3.1.1 Synchronisme

Au vue des résultats de l'expérience pilote, un remise en question m'est apparue nécessaire. Comme cela avait été suggéré dans les articles, le besoin de synchronisation dépend du stimulus, mais il avait aussi été dit que pour de la voix par exemple, la limite supérieure du décalage se situait au maximum à 250ms pour l'audio en retard, et de 125 pour des retards d'image. Est-ce que pour une relation audiovisuelle moins bien connue que la parole ces seuils seraient augmentés ? Cela me paraît manifeste et je vais tenter de le montrer.

3.1.2 Evolution temporelle

Si la question de la synchronie n'est pas l'actrice majeure de la fusion audiovisuelle, quelles lois, mises a part l'intuition, peut on utiliser pour fabriquer un ensemble audiovisuel convaincant.

Une des idées ayant décidé de la réalisation de ce stage consistait à essayer de comprendre les associations directes que l'on fait à la vue d'une séquence d'images. Une balle qui rebondit par exemple, si son mouvement n'est pas trop amorti et qu'elle semble dure, on associe immédiatement une succession de bruits de chocs assez courts de plus en plus rapprochés. Un exemple plus pratique repris des cours de Serge de Laubier présente l'image d'une forme suivant une trajectoire de spirale amortie à une vitesse constante.

Je vous laisse associer le son que vous voulez, et je suis sûr qu'il ressemblera à celui qu'imagine votre voisin.

Il est évident qu'ils ne seront pas identiques, mais ils auront au moins certaines caractéristiques morphologiques. Suivant les paramètres que l'on associe à la hauteur verticale

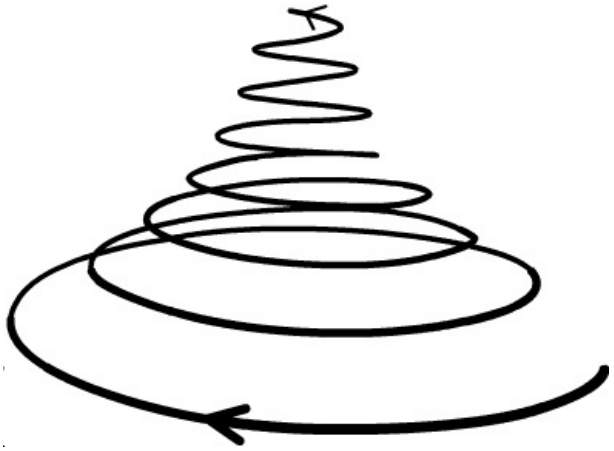

FIG. 3.1 – Spirale exponentiellement décroissante

(fréquence, intensité,...), à la profondeur du dessin (fréquence, intensité) on aura un son dont les paramètres varieront de la même façon que la forme.

La connaissances a priori de chacun sont influentes sur la qualité de l'ensemble, mais que l'on associe la dimension verticale à la fréquence croissante ou bien décroissante, tant que les caractéristiques morphodynamiques sont conservées, nous serons convaincus.

Aux vues de ces considérations, l'expérience de catégorisation entreprise dans la première tentative d'expérience semble quelque peu inutile.

Ces idées ont été suggérées par Noël château [CHAT], Michel Chion [CHION], citant la théorie de la Gestalt.

3.1.3 La théorie de la gestalt

La théorie de la gestalt est née à la fin du 19 ème siècle. C'est une théorie de psychologie dont les principaux acteurs sont Wertheimer, Köhler, Koffka et Lewin et s'opposant à la psychologie classique. Paul Guillaume, auteur de La Psychologie de la Forme [GUIL-79] et principal acteur de la Gestalt en France, décrit cette théorie comme la réponse aux insuffisance des théories associationiste et behavioriste. Une des idées fondatrices est celle selon laquelle l'association de corps crée un ensemble différent de la somme simple. Cette notion est contenue dans le mot même de Gestalt, malheureusement intraduisible, en allemand, *gestalten* signifie « mettre en forme, donner une structure signifiante ». Pour reprendre un exemple de Guillaume, une table sur laquelle sont disposés des livres a une signification très différente de la même table recouverte d'une nappe. Wikipédia [WIKI] présente un résumé des postulats Gestaltiste et des principales lois de la Gestalt :

Les postulats Gestaltiens

– Le monde, le processus perceptif et les processus neurophysiologiques sont isomorphes ; c'est à dire structurés de la même façon, ils se ressemblent dans leurs structures et les lois (d'une certaine façon).
– Il n'existe pas de perception isolée, la perception est initialement structurée.
– La perception consiste en une séparation de la figure sur le fond (vase de Rubin).
– Le tout est perçu avant les parties le formant.
– La structuration des formes ne se fait pas au hasard, mais selon certaines lois dites naturelles et qui s'imposent au sujet lorsqu'il perçoit.

Les principales lois de la Gestalt

– **La loi de la bonne forme :** loi principale dont les autres découlent : un ensemble de parties informe (comme des groupement aléatoire de points) tend à être perçu d'abord (automatiquement) comme une forme, cette forme se veut simple, symétrique, stable, en somme une bonne forme.
– **La loi de bonne continuité :** des points rapprochés tendent à représenter des formes lorsqu'ils sont perçus, nous les percevons d'abord dans une continuité, comme des prolongements les uns par rapport aux autres.
– **La loi de la proximité :** nous regroupons les points d'abord les plus proches les uns des autres.
– **La loi de similitude :** si la distance ne permet pas de regrouper les points, nous nous attacherons ensuite à repérer les plus similaires entre eux pour percevoir une forme.
– **La loi de destin commun :** des parties en mouvement ayant la même trajectoire sont perçues comme faisant partie de la même forme.
– **La loi de clôture :** une forme fermée est plus facilement identifiée comme une figure (ou comme une forme) qu'une forme ouverte.

Ces lois agissent en même temps et sont parfois contradictoires.

Nous nous attacherons à la loi de destin commun qui est l'unique loi permettant de décrire des déroulement temporels. Le seul test mené vérifiant la loi de destin commun pour l'audiovisuel est celle de la marionnette dont les mouvements de mâchoire sont corrélés ou non à la diffusion de la parole. Les résultats tendent à la vérifier, mais la loi de destin commun va plus loin que de la simple coordination, elle précise que si deux stimuli ont, d'un point de vue perceptif, la même trajectoire, alors ils ne forment qu'un seul objet. Les mouvements de machoire ne me paraissent pas correspondre à une réelle trajectoire temporelle, mais plutôt à ce que Michel Chion appelle le point de synchronisation, c'est à dire une ponctuation de l'audio par le visuel.

Nous allons essayer de rajouter à notre dispositif expérimental un test capable de mesurer l'effet de la trajectoire temporelle sur la fusion udiovisuelle.

3.2 le test

3.2.1 Le test temporel

L'importance de la synchronisation n'est pas a écarter du problème, l'expérience pilote a montré que l'auditeur percevait quelques fois des discordances sans pour autant être capable de régler le retard. Essayons de faciliter la tâche aux sujets en créant des séquences, mais plus courtes et en boucle afin que le sujet puisse s'habituer et utiliser l'anticipation citée par (max matthews).

Nous allons donc refaire la même expérience que précédemment, mais avec un autre type de stimulation. Max permet de générer des rampes de valeurs (de 0 à 127) qui vont diriger les synthèses à la place des capteurs, et mis en boucle, on obtient dans le temps un signal triangulaire. On peut par ailleurs ajouter une acélération à ces rampes, voici les trois types d'évolution que nous avons choisis :

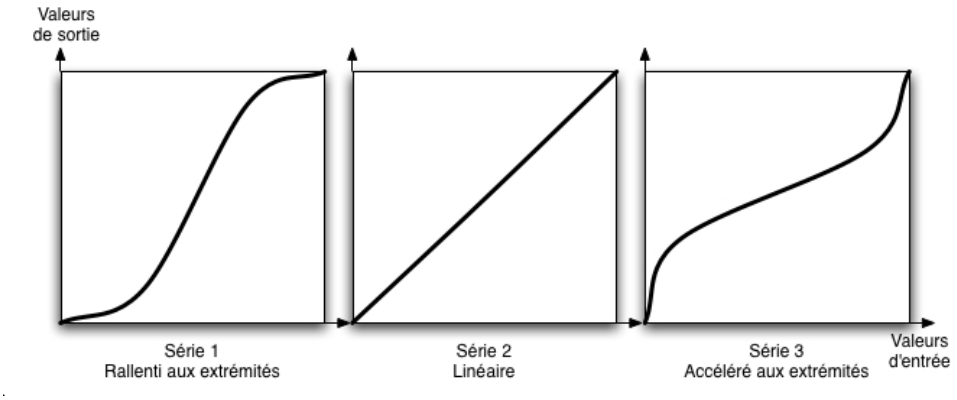

FIG. 3.2 – Les trois types d'évolution

Dans le cas de la série 1, le mouvement sera rallenti aux extrémités et donc fluide, pour la série 2, on a un mouvement linéaire, mais la rupture de pente crée une sensation légèrement percussive, et pour la série 3, le mouvement est très percussif.

3.2.2 Le test de modification de trajectoires

La première étape si l'on veut obtenir des trajectoires perceptivement identiques est de normaliser le trajectoires à l'aide des lois psychophysiques. L'intensité est perçue comme évoluant linéairement si elle suit l'échelle de sonie, la hauteur suit une loi logarithmique. Du point de vue visuel, la perception d'un objet se dirigeant vers nous à vitesse constante suit une courbe accélérée, mais si on se place face à un écran, la perception reste linéaire, et il en est de même pour les deux autres dimensions. La perception de la taille pour une sphère est en rapport avec le rayon, sur un écran, la perception est en deux dimensions et

on ne voit qu'un disque, nous opérons des rapports de surface, et un disque semble deux fois plus gros lorsque le rayon est multiplié par la racine de 2.

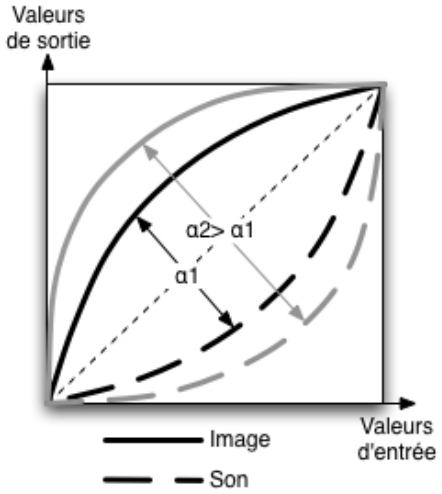

FIG. 3.3 – evolution en fonction de α

Mon idée pour obtenir une distinction graduelle entre les courbes a été de ramener les valeurs entre 0 et 1 et d'appliquer une fonction de puissance α pour l'image et $1/\alpha$ pour le son ou inversement. Les deux modalités ont donc leur trajectoire qui diffère selon le paramètre mais restent synchrones aux extrémités.

Nous avons donc testé des ensembles audiovisuels ne comprenant qu'un mouvement de chaque modalité, liaison par exemple de la fréquence de coupure du filtre d'un sample avec le déplacement de la boule sur l'axe vertical, et pour rendre le signal un peu moins monotone, nous avons appliqué un mouvement comportant une partie accélérée, une partie linéaire, et une autre ralentie. Cet ajout a été fait en accord avec le fait qu'une perception de stimulus strictement linéaire possède un caractère ennuyeux, ce qui risquait d'influencer le jugement des sujets (la remarque m'avait été faite lors de l'expérience pilote).

3.2.3 L'inclusion du geste

La troisième partie de l'expérience reprenait la précédente, mais en ajoutant le geste. Cette fois, le signal de base n'était pas délivré automatiquement, mais le sujet utilisait un des capteurs (en pratique, seul le capteur de pression a été utilisé). Il passait donc du statut d'auditeur à celui d'acteur.

Dans un quatrième temps, je laissais au sujets le loisir de jouer un peu avec les différents instruments en leur laissant configurer le mapping comme ils le voulaient.

3.3 Méthodes psychophysiques et logique des choix

3.3.1 La méthode des stimuli uniques

Le choix du protocole de mesure est très dépendant de l'hypothèse formulée. Nous aurions pu faire, comme les autres expériences de synchronie, une détection de seuil, mais je trouvais étonnant que la dégradation d'un ensemble audiovisuel soit brusque. Il fallait donc utiliser une échelle de partition. Ces échelles ont basées sur le fait que l'on peut découper le continuum subjectif en intervalles sensés. Nous avons plus précisément utilisé une échelle de cotation (rating scale), le sujet doit donner une note subjective à chaque stimulus d'une série. Ce type de méthode est décrit dans le Manuel pratique de psychophysique de Claude Bonnet [BONN], il consiste a présenter une série de stimuli auxquels ils doivent a chaque itération attribuer une note. Cette série doit être présentée sans référent, ni extrêmes et dans le cas d'un stimulus unique et de catégorie limitée (c'est a dire qu'il n'y a qu'un stimulus présenté à chaque fois), on doit utiliser la méthode des jugements absolus et utiliser une échelle.

3.3.2 L'échelle

Dans le cas des stimuli uniques, l'échelle doit être divisée en parties égales et disposer d'une cotation, chiffrée ou verbale.

Didier Delignieres [DELI] dans sa thèse de doctorat parle de l'échelle de catégories "RPE scale" étudiée par Borg en 1970 [BORG], une échelle numérique accompagnée d'expressions verbales, Borg à montré que le recours aux expressions verbales permet une évaluation plus précise que l'utilisation d'un étalonnage strictement numérique. Par ailleurs, le nombre de catégories ne doit être ni trop faible, ni trop élevé : et que l'effectif optimal se situe entre 5 et 7, ce qui n'empêche pas l'utilisation d'échelons intermédiaires, destinés à affiner l'évaluation. Cette échelle est utilisée dans le test type Mushra pour les jugements de l'encodage audionumérique.

Voici l'échelle que nous avons utilisé :

Elle comporte 6 catégories possédant chacune 3 valeurs pour préciser la réponse. Le choix de la notation sur 20 n'est pas non plus anodine, étant donné que nous y sommes assez habitués.

3.3.3 Le choix des stimuli

On peut lors d'une expérience de psychoacoustique présenter un grand nombre de fois le même stimulus, ce que j'ai cherché à éviter. En effet s'il avait toujours été le même, les sujets se seraient d'une part fatigués détériorant ainsi les dernières valeurs, et ils se seraient de plus habitués, et ce type de tâche demande suffisamment de concentration.

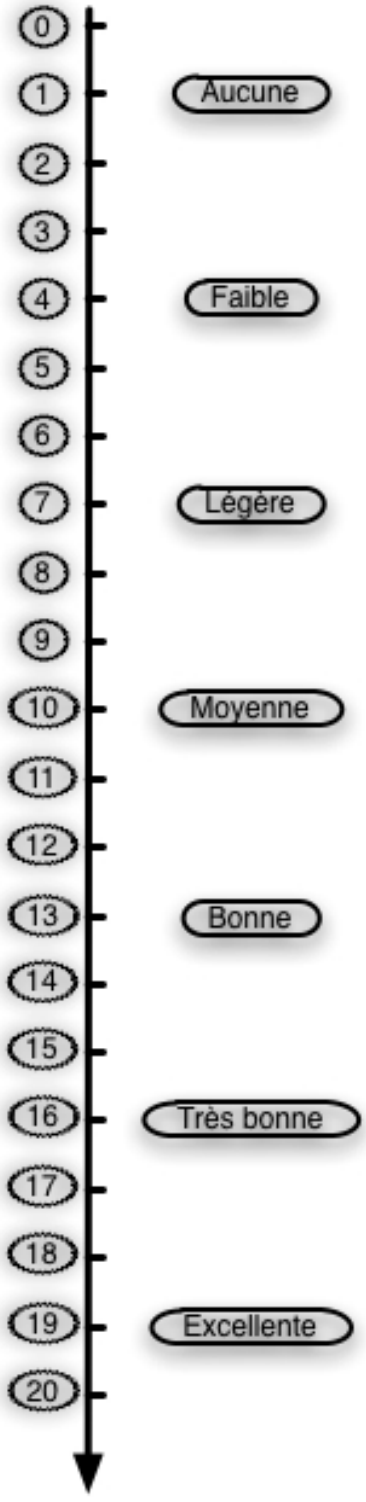

FIG. 3.4 – Echelle utilisée de type RPE scale

Nous disposions de quatres modalités visuelles : hauteur, taille position en x et en y de la boule (le nombre d'angles étant discontinu et la couleur étant une perception dont la linéarité d'évolution est difficile à obtenir) et de quatre modalités sonores : fréquence de coupure d'un filtre, hauteur de la fréquence fondamentale de deux sons différents (synthèse FM et filtres résonnants) et intensité de la synhèse FM. Cela nous mène à 16 possibilités, et en les répétant trois fois, on obtient 48 stimuli placés dans un ordre aléatoire mais évitant la succession de deux identiques, et l'ensemble passé trois fois pour les trois tests. Ce choix de diversité est discutable, et je le justifie par le fait que les sujets seront plus sensibles à certains stimulis qu'à d'autres, ce que nous voulons obtenir est plus une loi d'évolution et non pas une valeur précise d'un quelconque paramètre. De plus, l'apprentissage ainsi que l'ennui sont évités.

3.3.4 Les consignes

A leur arrivée dans la chambre d'écoute, les sujets étaient informés de la tâche, ils devaient juger de la qualité des ensembles audiovisuels qui allaient leur être soumis sachant que des signaux audio et vidéo seraient, et cela de manière aléatoire. Le sens de qualité était précisé, ils devaient juger de si ça « colle » ou pas et surtout ne pas tenir compte de la modalité visuelle ou sonore en jeu à chaque itération. Je précisais que par exemple si la balle se déplace verticalement, le fait que le son soit plus aigu ou plus grave lorsque la boule est en haut n'a strictement aucune importance et qu'il en est de même qu'une modalité vidéo ou une autre soit utilisée. L'accent était mis sur le caractère aléatoire et ils ne devaient pas chercher à trouver une quelconque logique.

J'ajoutais qu'ils devaient essayer de considérer un ensemble et de ne pas passer trop de temps sur chaque stimulus pour en donner un jugement relativement immédiat, une première impression.

Entre chaque série, je demandai de juger de la même chose et si possible d'utiliser l'échelle de la même façon.

3.4 Les Résultats

Nous avons effectué ces tests sur 20 sujets

3.4.1 Le test temporel

Voici le type de graphique obtenu a partir des résultats de la première expérience pour un seul sujet :

Ces quatre sujets ont eu des comportements différents.

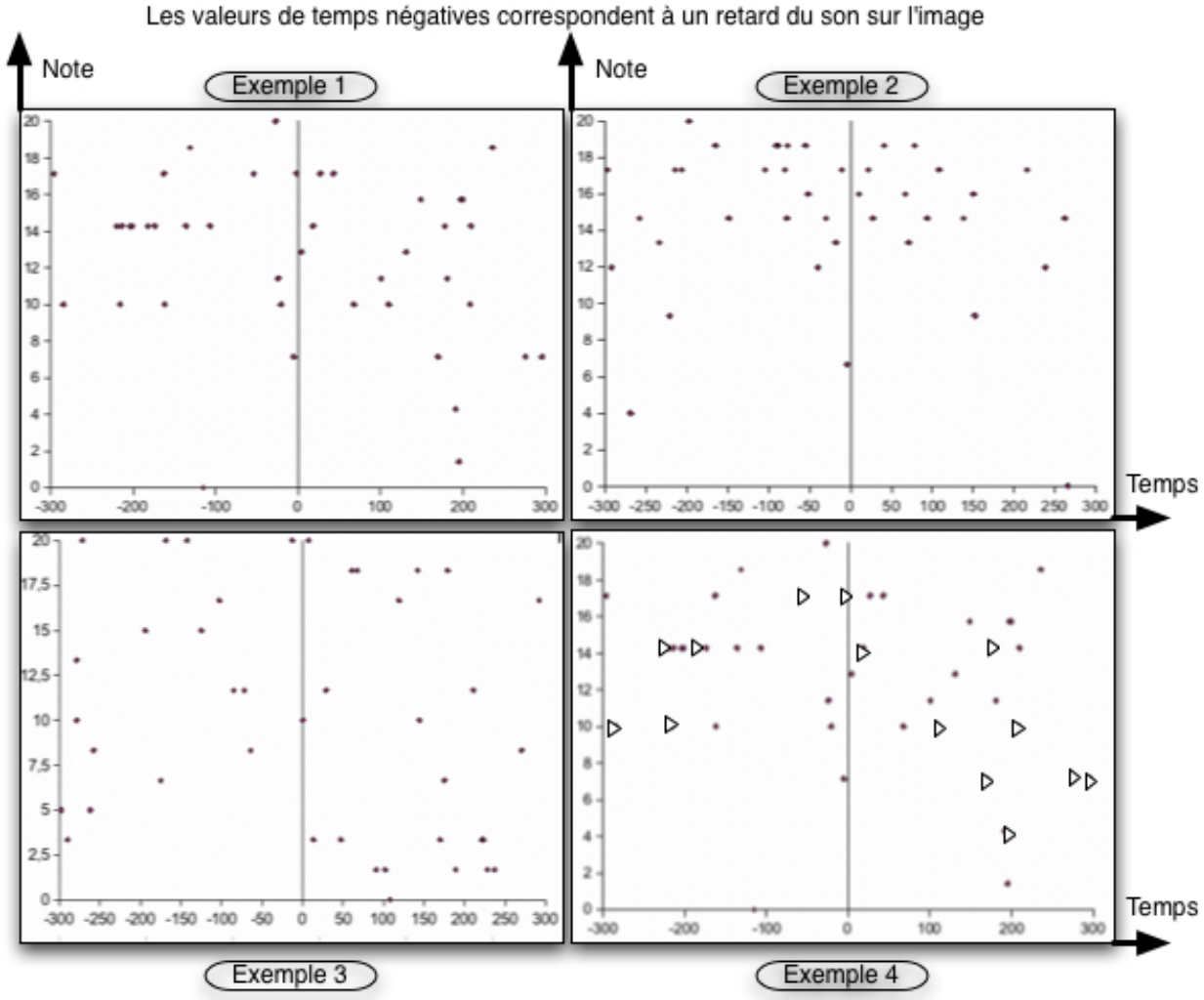

FIG. 3.5 – Exemples de résultats pour le test temporel

Le exemple 1 répond clairement au hasard ou ne suit pas la consigne, ce que j'ai pu vérifier en lisant les commentaires verbaux à la fin de l'expérience et malgré les préventions, cinq sujets ont noté uniquement leurs préférences, et donnent des réponses comme « j'ai toujours mal noté quand les aigus étaient en haut » ou bien « je n'ai pas du tout aimé ce son là ».

L'exemple 2 représente 3 sujets qui n'ont pas perçu ou peu de décalages, ce qui encore une fois s'est bien précisé lors de la verbalisation.

L'exemple 3 représente la plupart des sujets, ils perçoivent quelques fois des décalages, mais ce n'est pas systématique, leur verbalisation précise bien qu'ils les ont perçus et ont été très étonnés lorsque je leur ai présenté leurs résultats.

L'exemple 4 représente trois sujets qui ont bien remarqué les retards temporels, mais seulement lorsque le son arrive avant l'image, on constate donc une asymétrie, mais sans réelle précision ni systémtisme, d'autant plus que l'essentiel des valeurs significatives ont été obtenues avec les stimuli percussifs (petits triangles).

On peut donc dire que l'asynchronie possède bien ce caractère asymétrique, que leur perception dépend fortement du type de stimulus (percussivité, sémantique), mais surtout de la situation mentale de détection ou de jugement. Si l'on cherche a détecter un décallage temporel, on abaisse notre seuil, alors que si notre attention est focalisée sur la qualité de l'ensemble, on peut ne pas remarquer des décalages allant jusqu'à 300ms voire plus.

3.4.2 Le test évolutionnel

Les résultats pour la variation du paramètre sont les suivants, sachant qu'il variait entre 1 et 4, la déviation pouvait suivre au maximum une loi de racine quatrième. Un paramètre bêta tiré au hasard décidait qui du son ou de l'image suivrait la puissance supérieure à 1. Sur ce graphique nous avons mis un alfa négatif pour spécifier les cas ou la puissance supérieure à un affectait le son, puis remis les valeurs entre -3 et 3. Les échelles ont été normalisées pour effectuer des comparaisons, et les dix premières valeurs ont systématiquement été retirées des graphiques, étant considérées comme de l'adaptation à la tâche.

L'exemple 1 représente une nouvelle fois ceux qui jugent selon leurs préférences esthétiques.

Les nuages de points des exemples 2, 3 et 4 semblent vaguement tracer une courbe en cloche. Comme on ne peut objectivement dire qu'il s'agit d'une courbe, j'ai préféré délimiter les parties du plan sur lesquelles il n'y a pas de points. Les formes de cloches apparaissent plus ou moins selon les cas, mais on ne peut nier que pour un grand décalage de trajectoire les notes sont toujours faibles, et que lorsque les trajectoires sont identiques, elles sont toujours bonnes, et que cette dégradation est graduelle.

3.4.3 Le test évolutionnel avec Geste

Le test suivant était strictement le même que le précédent avec pour seule différence, le contrôle gestuel. Voici les résultats :

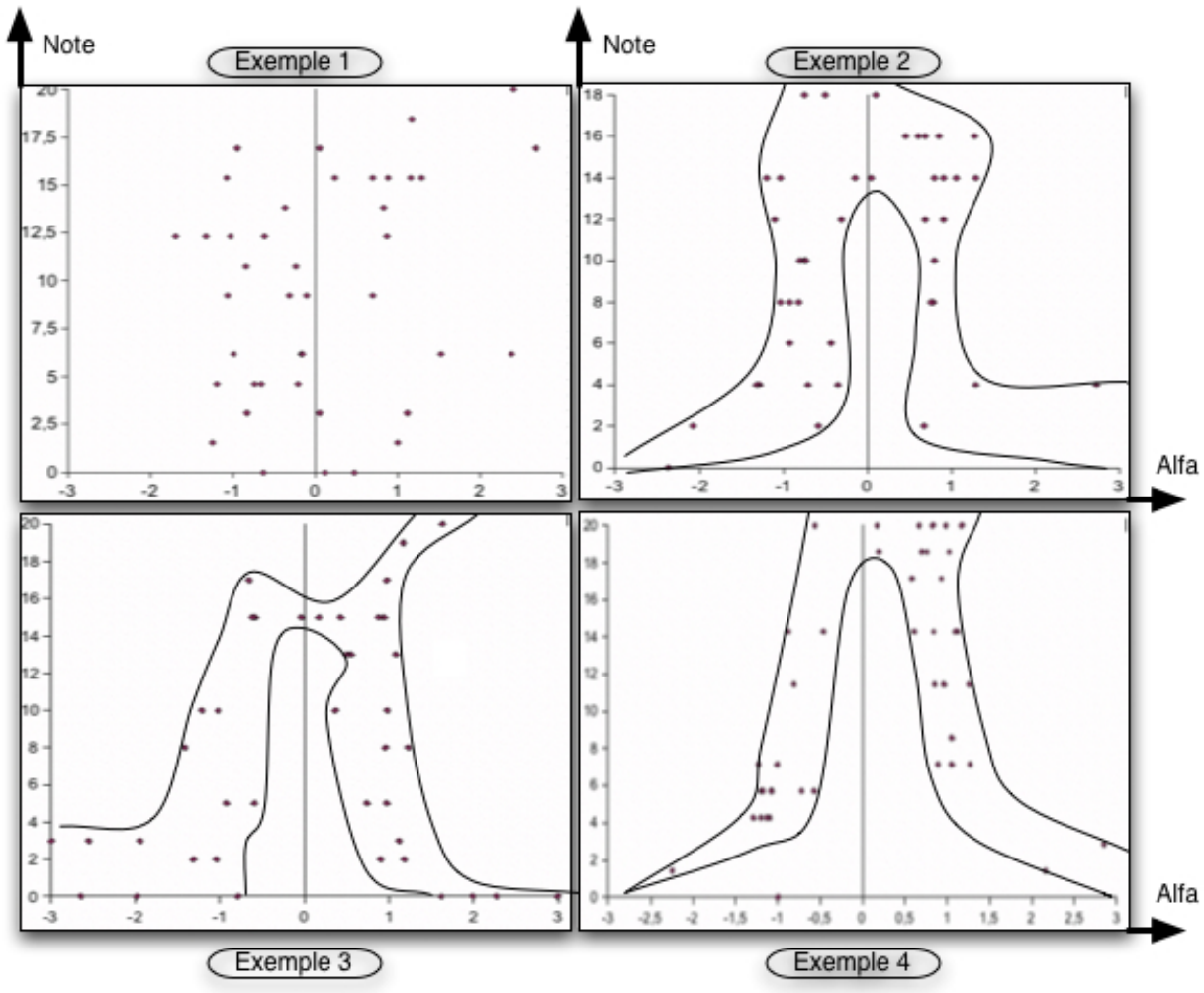

FIG. 3.6 – Exemples de résultats pour le test évolutionnel

Nous avons cette fois omi l'exemple des sujets n'ayant pas compris la consigne.

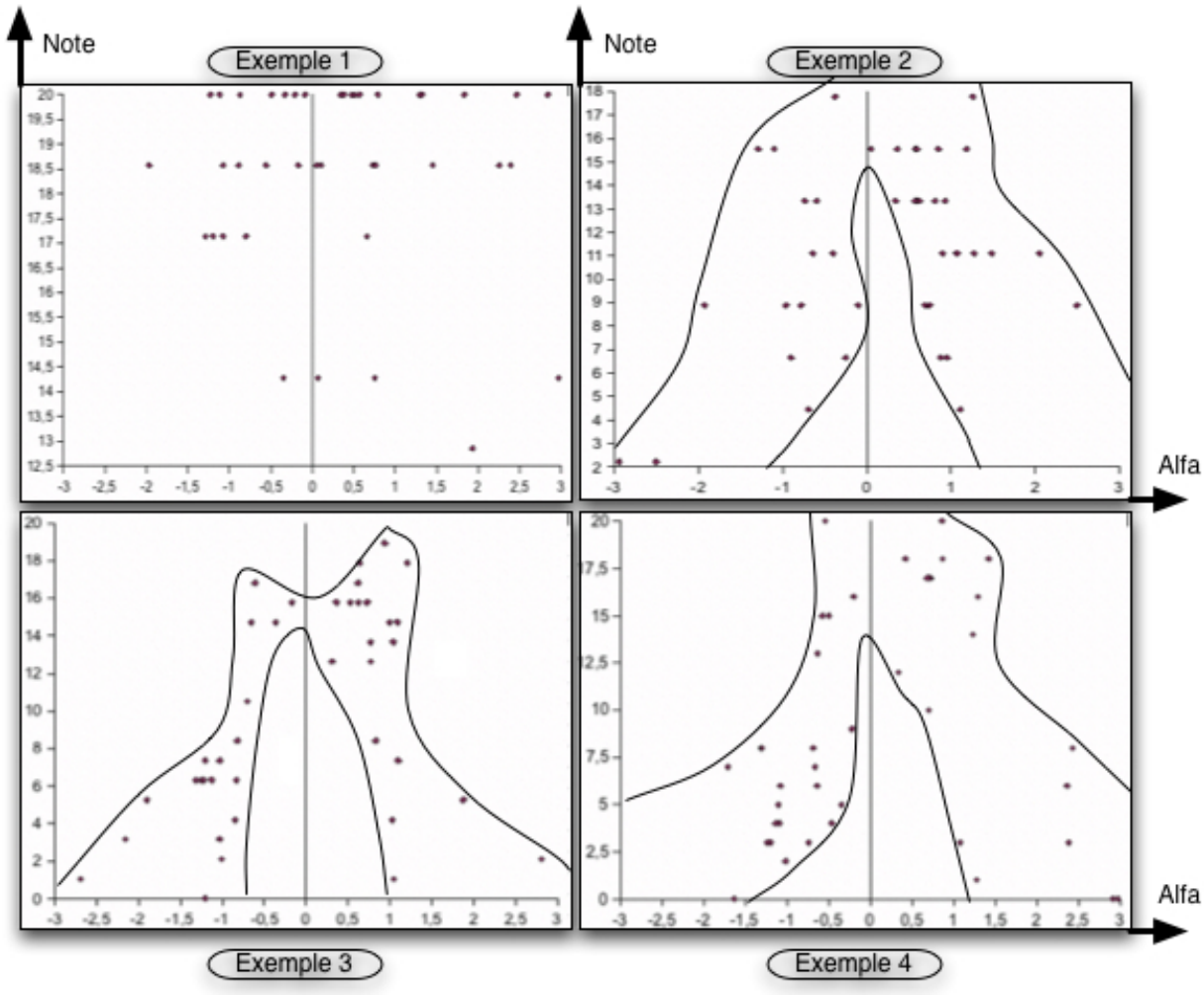

FIG. 3.7 – Exemples de résultats pour le test évolutionnel avec geste

On peut constater plusieurs effets de la présence du geste. En premier l'élargissement des nuages de points, ce qui denote une tolérance plus grande à des stimuli de trajectoires différentes. De plus, certaines fois, comme dans le cas de l'exemple 1, les notes sont extrêmement bonnes (elles ne descendent pas ici en dessous de 13. Dans des cas comme ceux des exemples 2, 3 et 4, en plus de l'élargissement, on observe une évolution des maximas par rapport au test sans geste, dans le cas des exemples 2 et 3, il a diminué, et dans celui de l'exemple 4, il a augmenté.

La présence de geste semble donc augmenter la tolérance vis a vis du décalage de trajectoire, parfois jusqu'à estimer qu'il n'est pas gênant. Le fait de contrôler l'objet renforce parfois la sensation de fusion, mais peut aussi la réduire, ceci étant peut être du aux aspirations

artistiques des sujets.

3.4.4 Jeux sur l'instrument

A la fin des trois tests, j'ai proposé aux sujets de jouer de cet instrument en utilisant tous les capteurs, tous n'y ont pas pris le même plaisir, mais leur laissant le choix du mapping, j'ai été étonné de la diversité des choix. Je m'attendais à ce que la tablette graphique soit liée à la la spatialisation et a la position de la boule en xy, et cela n'a été le cas que dans 50% des cas. Le changement du nombre de dimensions de la boule n'a presque jamais été utilisé, peut être est-ce du au fait que c'est une modalité discontinue.

J'ai souvent eu la remarque, après 20 à 30 minutes de jeux, « c'est vraiment marrant ton truc »... et pour avoir moi aussi joué avec, un tel instrument audiovisuel crée une espèce de relation instrumentale augmentée très agréable même malgré la simplicité de celui ci.

3.5 D'autres expériences envisageables

Aux vues des résultats de défrichages obtenus, il pourrait être intéressant de faire d'autres expériences pour les affiner.

Même si cela pourrait être ennuyeux pour un sujet, il pourrait être intéressant de choisir une modalité visuelle et une modalité sonore et de procéder aux mêmes tests. Il est évident que des préférences personnelles ont influencé les résultats.

Tester différents capteurs et donc différents gestes et déterminer leur influence sur la relation audiovisuelle, nous n'avons utilisé qu'un capteur de pression et donc un seul type de geste.

Nous pourrions imaginer beaucoup d'expériences, et l'outil développé permet d'en réaliser facilement un grand nombre.

Conclusion

La synchronie temporelle apparaît au travers des différent travaux comme une obsession et n'est peut être pas si indispensable, ce que souligne Levitin [LEVI] dans son article, en s'adressant aux concepteurs d'instruments électroniques. Certaines modalités comme le geste requièrent une précision temporelle, le problème de la latence lorsque l'on utilise un clavier nous le rappelle. Cependant, n'y a-t-il pas une habitude liée à ce phénomène ? L'orgue Cavaillé-Coll, qui est un Orgue pneumatique, de l'église Saint-Sulpice à Paris, et dont les soupapes se trouvent à 5 m du clavier, possède une latence de l'ordre de la demi seconde, et qui n'est pas constante selon le clavier utilisé. L'organiste apprend tout simplement à jouer avec ce retard et s'y habitue.

Nous sommes habitués à la parole, la tolérance est donc relativement faible. La précédence d'une modalité sur une autre alors que la situation est physiquement impossible pose cette fois un problème, celui de l'anticipation, si elle n'est plus possible, il y a incompréhension. En ce qui concerne les rapports image/son, le son ne peut physiquement préexister à l'image, ce qui explique les dissymétries observées.

Mais discutons plutôt de la fusion, qui vise à ne percevoir qu'un et un seul objet par deux modalités sensorielles différentes, ce que nous faisons d'ailleurs continuellement. On a pu remarquer que la théorie de la Gestalt avançait des lois très robustes valables dans des cas intramodaux comme intermodaux, même si le fait même d'être en situation multimodale pose des problèmes lors des expériences, ce qui d'ailleurs est explicable par le postulat gestaltiste du tout formant plus que la somme de ses parties. Il reste alors à préciser des lois d'évolution perceptives linéaires pour pouvoir appliquer les résultats à d'autres sons et images. Ce type d'approche a permis d'obtenir des résultats de fusion à un niveau basique, mais la part de sémantique propre à chacun participe beaucoup à cette fusion, c'est elle qui fera la différence entre des pas bruités au cinéma et une sphère qui viendrait rencontrer une paroi, des expériences de catégorisation pourraient compléter notre étude. Il reste de plus à mettre en évidence pourquoi deux sons, même dépourvus de sémantique, s'ils ont la même trajectoire, mais par exemple pas le même contenu harmonique fusionneront de manière inégale avec la même image.

Bibliographie

[ACROE] Association pour la Création et la Recherche sur les Outils d'Expression, http ://www-acroe.imag.fr/

[ALLA] L. G. Allan and A. B. Kristofferson, « Successiveness discrimination : Two models », Perception & Psychophysics, vol. 15 :1, pp. 37-46, 1974.

[BAUM-06] O. Baumann, MW. Greenlee, ÇÊNeural Correlates of Coherent Audiovisual Motion PerceptionÊÈ, Cereb Cortex. 2006 Aug 23 Institute for Experimental Psychology, University of Regensburg, 93053 Regensburg, Germany ; Department of Psychology, University of Oslo, 0317 Oslo, Norway

[BETE] P Bertelson et M. Radeau, « Cross modal bias and perceptual fusion with auditory-spatial discordance, » Percept. And Psychoph. ; 29, 578-584, 1981

[BORG] Borg, « Perceived exertion as an indicator of somatic stress. Scandinavian Journal of Rehabilitation Medecine », 2(2/3), 92-98. 1970

[BONN] Claude Bonnet, « Manuel pratique de psychophysique », Armand Colin – Collection U, 1986

[CAVE] C.Cavé, R. Ragot et M.Fano, « Perception of Sound-Image synchrony in cinématographic conditions », 4^{th} workshop on Rythm Perception and Production, F-Bourges, 1992

[CHAT-98] Noël château, « Les interactions sensorielles entre le son et l'image lors de la perception d'événements audiovisuels », Rencontres Musicales Pluridisciplinaires, Musiques en scène, 1998

[CHION-90] Michel CHION, « l'Audio-Vision son et image au cinéma » Nathan Université 1990

[COUR-05] Benoît courribet, « réflexions sur les relations musique/video et strategies de mapping pour maxmsp/jitter », cicm - université paris VIII MSH Paris Nord

[DELI] Didier Delignieres, « Approche psychophysique de la perception de la difficulté dans les tâches perceptivo-motrices », thèse 1993

[DICAC] http ://atilf.atilf.fr/academie9.htm

[DIXO] N. F. Dixon and L. Spitz, « The detection of auditory visual desynchrony », Perception, vol. 9 :6, 1980

[GANZ] L. Ganz, « Animal Behavior Processes », Journal of Experimental Psychology vol. 33, 1975.

[GUIL-79] Paul Guillaume, « La psychologie de la forme », Flammarion, 1979

[HENR-04] Cyrille Henry, « Physical modeling for pure data and real time interaction with an audio synthesis » proceeding, SMC 2004

[HUNT] A. Hunt, M.M. Wanderley & M. Paradis. « The importance of parameter mapping in electronic instrument design », Proceedings of the Conference on New Instruments for Musical Expression, 2002.

[JACK] C. V. Jackson, « Visual factors in auditory locations » Quart. J. Exp. Psychol., 5, 52-65, 1953

[JACQ-06]Christian Jacquemin, « An Eye for an Ear and an Ear for an Eye : Bidirectionnal Control in Virtual Multimedia Instrument Design », 2006

[JASK] P. Jaskowski, « Simple reaction time and perception of temporal order : Dissociations and hypotheses », Perceptual and Motor Skills, vol. 82 :3, Pt. 1, 1996

[LEVI] Daniel J. Levitin, Karon MacLean, Max Mathews and Lonny Chu « The Perception of Cross-Modal Simultaneity »

[PAIL-92] Jacques PAILLARD, CNRS-NBM, Marseille, Editions revueEPD, 1992

[LAUB-06] Serge de Laubier, « Meta-Instrument 3 : a look over 17 years of practice » NIME06, 2006

[LYON-98]"L'art en vidéo", festival cinéma/vidéo de Lyon Musée d'art contemporain, Lyon 1998

[MCGR] M. McGrath and Q. Summerfield, « Intermodal timing relations and audio-visual speech recognition by normal-hearing adults », Journal of the Acoustical Society of America, vol. 77, 1985.

[MONT] Nicolas Montgermont, « Modèles physiques particulaires en environnement temps-réel : Application au contrôle des paramètres de synthèse », Mémoire de DEA ATIAM 2005

[NATH-99] Chrisanthie NATHANAIL, « Influence des informations visuelles sur la perceptions auditive. Conséquences sur la caractérisation de la qualité acoustique des salles », Thèse, 1999

[OLEA] A. O'Learyand G. Rhodes, « cross modal effects on visual auditory object perception », Percept and Psychophys., 35, 565-569, 1984

[RADEA-87] M. Radeau et P Bertelson « Auditory-visual interaction and the timing of imputs. Thomas 1941 Revisited » Psychol. Res., 49, 17-22, 1987

[SCHAE] Pierre Schaeffer, « **Traité des objets musicaux », éditions** Le Seuil, 1977

[THOM] G. J. Thomas, « expérimental study of the influence of vision on sound localization », J. of EXP. Psychol., 28, 163-175, 1941

[THRU-73] W. R. Thurlow and C. E. Jack, « Certain determinants of the ventriloquism effect », percept and Mat. Skills, 36, 1171-1184, 1973

[VAND] Van de Par, Steven,Kohlrausch, Juola and James, « Some methodological aspects for measuring asynchrony detection in audio-visual stimuli »

[WESS] D. Wessel, « Timbre Space as a Musical Control Structure », Computer Music Journal 3,(2), 1979.

[WIKI] http ://fr.wikipedia.org/wiki/