Relation image/son : de l'illustration sonore à la fusion multi-modale

Hervé Zénouda

Relation image/son : de l'illustration sonore à la fusion multi-modale

Hervé Zénouda * (Paris le 15/10/2006)

* Musicien, concepteur/réalisateur/enseignant dans le domaine des hypermédias, doctorant (Université Paris 13).
Email : cargo18@free.fr
Site personnel : http://zenouda.free.fr

Résumé :
Nous essaierons, dans cet article, de regarder l'évolution, dans les hypermédias, de la pratique de l'illustration sonore généralisée par le cinéma. Si nous pensons que toute illustration implique une intention, un point de vue, une relecture du réel, ces choix dans les hypermédias n'impliquent plus uniquement les résultats médiatiques de surface mais se font à des niveaux bien plus profonds : sur les choix de structures, sur les primitives du modèle mis en œuvre.

Mots-clefs :
Sons et images numériques, interactivité, générativité, simulation, convergence, correspondance, fusion, mapping.

Abstract :
In this article, we will try, to look at the evolution, in hypermedias, of the sound illustration practice, generalized in cinema. If each illustration implies an intention, a point of view, a rereading of reality, those choices in hypermedias do not apply only to the media results of the surface anymore. They are as well involved in more abstract levels : choices of structures and primitives of the models in use.

Key words :
Digital sounds and pictures, interactivity, generativity, simulation, convergence, correspondance, fusion,, mapping.

Relation image/son : de l'illustration sonore à la fusion multi-modale

Nous essaierons, dans cet article, de regarder l'évolution, dans les hypermédias, de la pratique de l'illustration sonore généralisée par le cinéma. Si nous pensons que toute illustration implique une intention, un point de vue, une relecture du réel, ces choix dans les hypermédias n'impliquent plus uniquement les résultats médiatiques de surface mais se font à des niveaux bien plus profonds : sur les choix de structures, sur les primitives du modèle mis en œuvre.

Cette nouvelle situation implique des conditions de production et de réception particulières. Le son fusionne avec l'image dans le geste interactif et une nouvelle dimension de la perception apparaît, celle du processus en amont. La manipulation d'un hypermédia convoque ainsi à la fois l'activité et la contemplation, les effets perceptifs des images et des sons, les effets produits *entre* les différentes images et *entre* les différents sons et enfin la perception d'un processus *en amont* de ces images et de ces sons.

Rappel sur l 'illustration sonore au cinéma

Le son au cinéma est assujetti à l'image et ceci sous différents aspects. Historiquement, le cinéma a été muet avant d'être sonore et de nombreux interprètes (musiciens, acteurs, bruiteurs), dans la salle, sonorisaient en direct les images projetées. On a pris ainsi l'habitude que ce soit le son qui s'adapte à l'image et non l'inverse. De fait, chaque soir, et selon l'inspiration du moment, cette enveloppe sonore pouvait changer dans le contenu, dans le choix des éléments à sonoriser et dans le calage temporel. Une fois que le son a été enregistré sur le même support que l'image, cet aspect d'improvisation a disparu et les relations complexes entre les deux représentations ont été de la responsabilité du réalisateur. Aujourd'hui, dans le processus de production d'un film, à l'exception des sons et dialogues enregistrés pendant le tournage, de nombreux éléments de la « bande sonore » [1] sont travaillés dans une période de post-production, c'est-à-dire à un moment où les images sont entièrement montées. Ceci concerne la musique, les bruitages, les doublages des voix. Et même si on pouvait trouver de nombreux exemples de cinéastes qui prennent en compte des éléments sonores dans des étapes antérieures de production (à l'écriture du scénario, au tournage ou au montage), la position en fin de parcours de fabrication incite à ce que ce soit le visuel qui soit sonorisé et non le sonore illustré.

Ceci n'empêche pas, de nombreux effets de fusion image/son d'apparaître à la réception. Michel Chion[2] a développé les notions de *valeur ajoutée* et de *synchrèse* pour décrire ces effets perceptifs. La valeur ajoutée est la « valeur expressive et informative dont un son enrichit une image donnée, jusqu'à donner à croire que cette information supplémentaire se dégage 'naturellement' de ce que l'on voit et est déjà contenue dans l'image seule. Et jusqu'à donner l'impression que le son redouble un sens qu'en réalité il amène et crée, soit de toutes pièces, soit par la différence même d'avec ce qu'on voit … »[3]. Ce phénomène est perceptible surtout dans le cadre de la synchronisation son-image qui permet de nouer une relation immédiate entre ce que

[1] Michel Chion récuse, à raison, l'idée d'une « bande son » analysable indépendamment de l'image.
[2] Michel Chion : l' *Audiovision* (Nathan, Paris 1990) et *Un art sonore, le cinéma* (Cahier du cinéma, Paris 2003).
[3] Michel Chion, l' *Audiovision*, Nathan, Paris 1990, p. 8.

l'on voit et ce que l'on entend. Pour souligner cette fusion perceptive, Chion crée le terme de *synchrèse*, mot-valise concaténant les termes synchronisme et synthèse. Cette fusion perceptive ne met pourtant pas l'image et le son sur le même plan d'égalité, mais réaffirme à nouveau la primauté de l'image sur le son : « La valeur ajoutée est réciproque : si le son fait voir l'image différemment de ce que cette image montre sans lui, de son côté, l'image fait entendre le son autrement que si celui-ci retentissait dans le noir. Cependant, à travers ce double aller-retour, l'écran reste le principal support de cette perception. Le son transformé par l'image qu'il influence re-projette finalement sur celle-ci le produit de leurs influences mutuelles»[4]. De même, la position spatiale du son ainsi que les espaces fictionnels créés par lui (les différents champs : *In*, *Hors* et *Off*) sont dépendants de sa relation à l'image. La synchronisation fait coller la dimension sonore au visuel et permet d'associer le mouvement imaginaire du son à celui, réel, de l'image. De même, chaque changement de cadre peut transformer le statut d'un son perçu comme *Off* ou *Hors* en son *In*.

D'autre part, le son au cinéma permet de jouer sur les dimensions de l'image, c'est-à-dire sur sa *profondeur*, sa *largeur*, sa *durée*. Le jeu sur la *profondeur* est obtenu en mixant et superposant différents plans sonores, le son ajoute ainsi une profondeur, une épaisseur à l'image. Le jeu sur la *largeur* est obtenu grâce au hors champ qui permet de pousser les limites spatiales du cadre. Enfin, le jeu sur la *durée* permet de relier entre eux plusieurs plans, soit en utilisant le son comme raccord d'un plan à l'autre, soit en les unifiant dans un même flux sonore. Enfin, du point de vue du sens, le son peut être dans une relation de redondance, d'opposition, de complémentarité ou d'indifférence avec l'image. Signalons que la relation de redondance souvent décriée n'est pas si triviale qu'il y paraît. En effet, une même information exprimée dans des canaux sensoriels différents est toujours porteuse d'éléments complémentaires. La redondance parfaite n'existe pas entre le son et l'image.

Comment ces éléments d'illustration sonore sont-ils pris en compte dans les hypermédias ? La relation entre l'image et le son nous semble évoluer sensiblement dans la situation interactive.

Images et sons dans les hypermédias

Les hypermédias proposent une expérience participative où l'action et la perception s'entremêlent étroitement. Le corps, et particulièrement le geste, fait ainsi irruption dans nos relations avec les images et les sons. L'interactivité, dimension centrale des hypermédias, se décline dans différentes situations. Elle peut s'appliquer à des objets finis, à des objets générés ou à des êtres ou des mondes simulés. Chacun de ces cas implique des relations particulières entre les deux représentations ainsi que des effets perceptifs particuliers. Ces situations aussi bien techniques que communicationnelles peuvent se distinguer par l'introduction de la notion de *causalité*. La première situation entraine que la cause des résultats médiatiques pointe vers l'utilisateur. Dans la seconde, la cause est à trouver dans les règles de production du générateur. Enfin dans la troisième, la cause pointe vers le monde ou l'être simulé, source de toutes les manifestations médiatiques.

[4] *Idem*, p. 22.

Ceci n'est pas sans effets sur la table de correspondance (le mapping) entre les différents médias. De ce mapping, qui est au centre des choix créatifs des concepteurs d'œuvres interactives, découle les possibilités manipulatoires de l'œuvre.
Dans le premier cas, les concepteurs tissent des liens entre des images et des sons préconçus. Les caractéristiques d'un domaine sont mises en relation avec celles de l'autre domaine avec comme intermédiaire le geste qui peut lui-même être augmenté, diminué ou laissé tel quel. Les liens tissés peuvent consister en des relations de plusieurs natures : une relation de un à un fait correspondre un paramètre d'une dimension à un paramètre de l'autre, mais cette relation peut aussi être de un à plusieurs ou de plusieurs à un. Cette interaction sur des objets finis est dans la continuité des *correspondances sensorielles* telles qu'elles ont été expérimentées au début du siècle par des artistes comme Kandinsky. Avec une interactivité sur des objets générés, la correspondance se fait plutôt par le biais d'un métalangage de communication entre les différents générateurs de médias. Chaque générateur dégage des structures (durées, ordre de thèmes …) pouvant être utiles aux autres générateurs. Le mapping se fait ici via une étape de traduction dans une structure langagière. Les correspondances ne s'effectuent plus sur les caractéristiques sensibles des images et des sons mais sur leurs modes de production. Ce type de mapping s'intéresse ainsi à des *correspondances structurelles*. Enfin, dans la troisième situation, un niveau d'abstraction est franchi. Il ne s'agit plus de coordonner plusieurs générateurs obéissant à des règles spécifiques mais d'avoir un seul objet programmé (un être ou un monde simulé) avec lequel on communique et qui s'exprime par des générateurs de médias différents. On pourrait parler alors d'un langage en amont des médias, d'un langage « a-média ». Ce type de mapping est spécifique au numérique et correspond à une situation de *fusion des modalités* dans du langage.

De nouveaux effets perceptifs

Ces situations de production particulières ont un impact sur la réception avec des effets perceptifs spécifiques. Les hypermédias ancrent le son à l'image de manière plus forte qu'au cinéma en accentuant les synchronisations intempestives dues à la manipulation interactive. Cet apparent asservissement du son à l'image ne doit pas masquer la transformation de cette même image au contact rapproché du son. Ainsi dans les œuvres qui proposent des jeux sonores et visuels riches, le rapport entre le son et l'image s'inverse souvent. En manipulant une image, on produit du son. Après quelques instants, l'intérêt sonore et l'intérêt visuel viennent s'entremêler de manière à ne plus savoir si c'est le son ou l'image qui dictent les mouvements du geste. Les effets de synchronisations se multiplient dans une synchrèse généralisée liant images, sons et gestes. De plus, cette synchronisation s'effectue aussi dans le sens inverse, de la substance audiovisuelle vers l'interacteur. Dans le cas de la manipulation d'objets temporels comme l'image animée ou le son, l'interacteur manipule ces objets mais se synchronise aussi sur eux en adaptant son geste à leurs propres temporalités. Il s'agit ici d'un effet de *syntonie*[5] qui a été étudié, par exemple, par Jean Louis Weissberg[6].

[5] Syntonie : « égalité de fréquence des oscillations libres (de deux ou plusieurs circuits) […] circuits en syntonie : accordés sur la même longueur d'ondes. Syntonisation : […] Réglage de résonance qui assure le rendement maximum. » (le Petit Robert).

[6] Jean Louis Weissberg , *Corps à corps – à propos de la morsure*, http://hypermedia.univ-paris8.fr/seminaires/semaction/seminaires/txt01-02/fs-02.htm, Mars 2002.

D'autre part, en associant plusieurs sons possibles à une image, les hypermédias matérialisent l'axe paradigmatique développant ainsi un déroulement temporel particulier lié à l'interaction. Selon les règles d'association de ces sons (par calcul, tirés aléatoirement, par famille de sons), se développent différentes significations et une perception globale de l'ensemble de ces sons dans leurs rapports à l'image. Le sens se construit ainsi autant *entre* les sons et les images d'un même ensemble que dans leurs significations spécifiques.

Enfin, dans les systèmes interactifs qui mettent en jeu des fonctions génératives ou des situations de communication avec des entités ou des mondes simulés, la cause des émergences médiatiques (les règles de production dans le premier cas, le modèle de simulation dans le second) incite l'interacteur à dépasser les effets perceptifs de ces images et sons générés pour percevoir le processus *en amont*. Un processus qui n'est pas abstrait mais qui s'incarne dans la substance audiovisuelle imbriquant ainsi les différents niveaux de perception. C'est la spécificité des hypermédias d'associer étroitement ces aspects du sensible et du calcul[7] dans une même expérience. On peut dire que, de même que le défilement linéaire des photogrammes au cinéma fait apparaître le mouvement, la manipulation interactive fait apparaître le processus.

Quelques exemples

Une brève description de quelques exemples d'œuvres interactives nous permettra d'éclairer les nouvelles relations entre les images et les sons qu'instaurent les hypermédias.

Legato, *Cellos* et *Moon tribe*[8] sont trois réalisations qui mettent en scène des personnages filaires de danseurs. À partir de ces mêmes éléments graphiques, chacune de ces trois réalisations exploite un aspect musical bien particulier : *Legato* met en place un mixage de lignes mélodiques sur un thème en boucle, *Cellos* ordonne différentes mélodies préétablies, *Moon tribe* permet de synchroniser des boucles rythmiques. Ces trois propositions de Nicolas Clauss s'inscrivent dans la catégorie des jouets sonores[9]. En effet, la structure même de l'interaction est calquée ou transposée de structures musicales comme l'harmonie, le contrepoint, l'ordonnancement de mélodie, la synchronisation de rythmes. L'image n'est pas ici une représentation du sonore, mais possède sa propre cohérence esthétique installant ainsi des effets de correspondance arbitraires avec le son. De même, chacun des modes, visuel et sonore, possède son propre temps ; le temps de l'animation graphique et le temps musical ne se recouvrant pas toujours parfaitement produit ainsi des décalages des boucles temporelles visuelles et sonores. Le geste ne fusionne pas les deux modes mais les coordonnent. Il est le point de rencontre, la

[7] Lev Manovitch, dans son livre *The language of new media* (MIT Press, 2001) utilise les expressions de « *couche culturelle* » et de *« couche computationnelle* » pour désigner ces deux mondes. Pour lui, les nouveaux médias devenus programmables *(« cinématographique en surface, digital dans leur structure et computationnel dans leur logique* »), appellent ainsi une nouvelle approche qui va chercher ses modèles dans le domaine de l'informatique.

[8] http://www.flyingpuppet.com.

[9] Un jouet sonore est un dispositif qui met en œuvre une interface graphique permettant de manipuler certains paramètres du sonore et du musical. Le dispositif permet ainsi un certain espace de jeu. À la différence des instruments de musique qui nécessitent un certain apprentissage, ces règles de jeu sont suffisamment fortes pour que le résultat sonore garde toujours une certaine cohérence.

frontière entre ces deux mondes. Si cette relation provoque une fusion sensorielle, elle ne peut être que fugitive, non-univoque et soumise aux interprétations variant assez fortement d'un utilisateur à l'autre. Elle n'est pas directement inscrite dans le processus technique mais plutôt réalisée, au détour d'un geste intempestif, comme un plus d'appropriation de l'utilisateur.

La série interactive appelée *Zoo*[10] propose un bestiaire interactif où chaque animal traité possède des propriétés interactives particulières. Chacun des tableaux propose quelques principes communs simples : un animal traverse l'écran, l'interactant peut alors le saisir, le manipuler, le disloquer, bref découvrir ses propriétés cachées. La manipulation du personnage déclenche des modifications sonores et visuelles qui peuvent modifier radicalement l'atmosphère initiale. Le son participe à l'identité du personnage au même titre que ses autres attributs (visuels ou interactifs). L'évènement sonore déclenché va ainsi jouer sur la manière dont l'utilisateur manipulera le personnage qui elle-même sera influencée par les modifications visuelles et sonores ainsi générées. Cet évènement sonore peut être décrit comme une mélodie interne liée au personnage interactif et est constitutive du dialogue qui se noue avec l'utilisateur puisque directement modifié par ce jeu. Si dans les tableaux de Nicolas Clauss, on ordonnançait et mixait des éléments sonores entre eux via l'image sans pouvoir modifier leurs intégrités, ici l'interactivité modifiera directement la matière audiovisuelle. Des modifications des paramètres sonores vont être ainsi liées aux modifications visuelles induites par le geste (modification du volume, de la vitesse de lecture et de hauteur de notes, de la position spatiale).

Le CD-rom dédié à Léopold Senghor[11] propose un système de musique générative qui permet des variations infinies pour un même écran. Le sommaire est sonorisé par une musique dont le mixage est modifié par le déplacement du curseur et le survol des icônes dédiés aux différents chapitres. Une base de données sonore permet d'associer de nombreux sons à une même interaction[12].

Développeur de palettes graphiques qu'il sonorise en générant en temps réel des sons de synthèse, Golan Levin[13], associe des paramètres du geste comme la direction et la vitesse du mouvement ou la pression du crayon électronique à des paramètres sonores comme le timbre, la hauteur, le panoramique et, graphiques comme la couleur, l'épaisseur de trait ou la direction. Chez Golan Levin, les notions d'interactivité et de générativité sont étroitement liées : l'image comme le son sont générés en temps réel, à partir du geste de l'utilisateur. Avec *Yellowtrail,* le comportement de l'animation dépend strictement de la forme dessinée et de la vitesse du mouvement de l'utilisateur. Levin utilise ensuite ces graphismes comme spectrogramme inversé[14] pour générer le son. Un spectrogramme est la représentation graphique d'un son défini par ses informations de temps et d'amplitude. La technique de spectrogramme inversé permet de faire le chemin inverse, partir d'une image pour générer un son[15]. Dans *Loom*, les paramètres du

[10] Frédéric Durieu (programmation, image), Jean Jacques Birgé (musique), www.lecielestbleu.com, 2001-2002
[11] *Senghor*, Jériko, Paris 1999.
[12] *La musique du cédérom Léopold Sédar-Senghor*, http://perso.orange.fr/roland.cahen/
[13] http://acg.media.mit.edu/people/golan/

[14] « Pattern playback ».
[15] Voir des logiciels du type de *Metasynth*.

geste sont appliqués aux paramètres de la synthèse en modulation de fréquence[16]. Ainsi, la vitesse du trait contrôle l'amplitude, la pression du crayon influe sur le vibrato du son ainsi que son amplitude, les courbes du trait influent sur la modulation. Chaque objet graphique animé a sa propre tête de lecture et contient donc son propre temps permettant de créer de riches effets de polyrythmie. Avec *Aurora*, l'interactant dessine des nuages animés de couleur qui obéissent au geste. Ces nuages sont constitués de centaines de filaments de couleur qui sont associés aux nombreux grains sonores générés par une synthèse de type granulaire. Si dans *Loom*, le contrôle de la synthèse en modulation de fréquence nécessitait quelques dizaines de paramètres, dans *Aurora*, la synthèse granulaire en nécessite plusieurs milliers ! Il s'agit donc de faire correspondre les données des nuages visuels à la multitude de données que nécessite cette synthèse. De plus, cette mise en relation doit être signifiante au niveau perceptif pour l'utilisateur et l'on sait que distribuer directement les données du domaine visuel au domaine du sonore, ne donne pas souvent le résultat escompté. Pour créer un espace de traduction entre ces deux ensembles de données, Levin fait appel à des techniques de distribution statistique pour regrouper, augmenter ou réduire le nombre d'informations et les associer de manière rationnelle et signifiante.

Antoine Schmitt[17] est un artiste numérique qui propose des œuvres centrées sur une communication avec des entités dotées de comportements. L'activité de programmation au cœur de son travail met en avant la notion de processus au détriment d'un travail plus traditionnel sur les images et les sons. Ainsi, Antoine Schmitt va utiliser des moyens visuels et sonores volontairement réduits pour mieux concentrer ses œuvres sur les causes de ces émergences médiatiques : le comportement de l'entité avec laquelle nous entrons en communication. Chaque œuvre est donc tout d'abord à comprendre comme une expérience communicationnelle entre un interactant et une entité douée de comportements et exprimant sa « présence » par le biais d'images et de sons générés entièrement par le programme. Cette perception du comportement est principalement rendue par le mouvement et l'interrogation sur cette cause peut se résumer à « pourquoi ça bouge ? ». La phase de conception/réalisation, elle, est abordée comme une approche de simulation qui cherche à expérimenter des lois physiques imaginaires[18].

Conclusion

Au travers de ces quelques exemples, on peut percevoir ces nouvelles modalités techniques et perceptives énoncées plus haut. Avec le numérique, on assiste à un mouvement manifeste de convergence de l'image et du son. Elle s'exprime tant du côté des techniques de production (avec un codage commun et des représentations de haut niveau communes) que du côté de la perception avec la dimension interactive. Ce rapprochement aboutit à une fusion des deux modalités dans un nouvel objet audio-visuel.

Dans un premier regard, on pourrait penser que l'interactivité asservit encore un peu plus le son à l'image mais dans le même temps, le son gagne en importance dans cette fusion. Le son devient un élément constitutif de ce nouvel objet au même titre que l'image et n'est plus dans un rapport d'illustration avec elle. Le programme en

[16] Popularisée par le DX7 dans les années 1980.

[17] http://www.gratin.org/as/.

[18] *Le travail du temps : programmer « un mode d'être, un entretien avec Antoine Schmitt »,* Samuel Bianchini, http://www.gratin.org/as/txts/index.html.

amont des émergences médiatiques unifie les modes de production des images et des sons et leur donne une cohérence structurelle commune.
L'intention de l'auteur ne s'exprime plus dans le choix des représentations de surface mais bien dans le choix des modèles mis en œuvres ainsi que dans le choix de tables de correspondances entre l'image, le son et le geste. Une opacité nouvelle apparaît ainsi. L'intention est cachée dans le programme et se révèle au cours de nombreuses manipulations qui seules permettent d'entrevoir une partie des variations possibles permises par le système.
Face à cette ouverture de la forme, de nouveaux outils d'analyse doivent être mis en place qui prennent en compte l'activité de l'utilisateur et la variabilité des émergences médiatiques.
On rejoint ainsi les problématiques de l'oeuvre processus versus l'œuvre objet posées et théorisées par John Cage[19] dans les années 50.

[19] John Cage, *Pour les oiseaux*, Belfond, Paris 1976.