LA SONIFICATION DE SÉQUENCES D'IMAGES À DES FINS MUSICALES

Jean-Marc Pelletier
Keio University
Graduate School of Media and Governance
5322 Endo, Fujisawa, Kanagawa, Japon
jovan@sfc.keio.ac.jp

RÉSUMÉ

Cet article décrit une méthode pour la sonification du mouvement dans des séquences d'images à des fins d'expression musicale. Le système veut permettre au compositeur de chercher des idées musicales dans le plus grand nombre possible de structures visuelles. C'est pourquoi une approche de bas niveau, basée sur un ensemble de vecteurs décrivant le mouvement d'un nombre variable de points saillants dans l'image a été choisie. Chaque vecteur de mouvement est associé à un élément sonore, soit un grain dans le cas de synthèse granulaire, ou un partiel pour la synthèse additive. Grâce à certains effets psychologiques de *Gestalt*, cette approche permet la création d'objets sonores perçus correspondant à des objets visuels, sans toutefois avoir à identifier ces objets préalablement.

1. INTRODUCTION

Dans cet article, une approche de la sonification de séquences d'images à des fins musicales sera présentée. L'expression « séquence d'image » est voulu dans un sens très large. Une image est ici tout champ bidimensionnel représentant quelque chose. Une image peut donc provenir d'une caméra, ou de n'importe quelle autre technique d'imagerie, être un plan d'une séquence d'animation, être générée en temps réel ou non, etc. La séquence d'image doit représenter une progression temporelle car le mouvement qui s'y trouve représenté servira de base à la sonification.

La sonification est définie comme « la transformation de relations entre des données en relations perçues dans un signal acoustique afin de faciliter la communication ou l'interprétation»[11]. Cette définition, cependant, n'est pas tout à fait satisfaisante dans un contexte artistique car les concepts de communication et d'interprétation y sont mal définis. La sonification d'image sera donc ici définie comme une traduction d'ordre iconique du visuel vers le sonore, c'est-à-dire que la transposition s'effectue dans le domaine morphologique (écoute/vision réduite) plutôt que sémantique ou causal [3].

2. PRÉCÉDENTS

2.1. Aperçu historique

La sonification d'image a été rendue possible d'abord par l'arrivé de la piste son optique au cinéma. Comme une piste son dite à densité fixe représente l'onde sonore par un tracé graphique, il devient facilement envisageable de synthétiser de nouveaux sons en y photographiant diverses formes. Déjà, en 1929, à Léningrad, lors de la réalisation du premier film sonore soviétique (*Plan Vélikikh Rabot*, d'Abram Room), l'animateur Mikhaïl Tsékhanovski, en observant pour la première fois une piste son, proposa aux compositeurs Arséni Avraamov et Ievgéni Cholpo de chercher des musiques anciennes en photographiant des ornements égyptiens ou grecs [1]. Bien qu'Avraamov déçut Tsekhanovski en expliquant qu'un motif strictement périodique ne pourrait jamais produire de mélodie, il expérimenta lui-même la production de nouveaux timbres à partir de formes géométriques et l'inventeur Boris Iankovski, par curiosité, obtint un timbre rappelant celui du violoncelle à partir de profils en silhouette de ses collègues [7]. Vers 1932, l'animateur Allemand Oscar Fischinger travaillant alors sur son film expérimental *Tönende Ornamente* affirma avoir entendu un son rappelant un sifflement en photographiant l'image d'un motif égyptien de serpent sur une piste son [5]. Par la suite, en 1933, László Moholy-Nagy réalisa un film, *Tönendes ABC*, dans lequel autant la trame sonore que visuelle étaient composées de formes trouvées, tel des lettres ou des empreintes digitales, de manière à « montrer ce qu'on entend » [9].

Dans les années 1970, la vidéo a permit la sonification d'image en temps réel. Le synthétiseur *DIMI-O* (1970) d'Erkki Kurenniemi permettait de contrôler la génération de notes à partir d'un signal vidéo qui pouvait provenir d'une caméra [14]. En 1974, les artistes David Behrman, Bob Diamond et Robert Watts réalisèrent une pièce intitulée *Cloud Music* dans laquelle, à l'aide d'un système d'analyse vidéo, ils purent transformer les changements d'intensité de six points d'une image en signaux de contrôle analogiques, ce qui leur permit la sonification des mouvements de nuages, captés par une caméra [20]. Yasunao Tone, en 1976, pour sa pièce *Voice and Phenomena*, contrôla un synthétiseur en projetant des

images de caractères chinois sur un écran sur lequel il avait simplement installé quelques cellules photosensibles [4].

Par la suite et jusqu'à ce jour, les ordinateurs ont permis l'application de techniques de sonification plus poussées. À partir de 1983, l'artiste David Rokeby, dans des pièces comme *Reflexions*, *Body Language*, et *Very Nervous System* (1986-1990) [18] utilisa le mouvement capté par des caméras pour générer des événements sonores – différentes régions de l'image correspondant à différents sons. Yasunao Tone réalisa d'autres pièces utilisant des idéogrammes chinois, cette fois-ci utilisant des techniques numériques. Par exemple, dans *Musica Iconologos* (1992), les données d'une image numérisée sont tout simplement ré-interprétées en tant que courts fragments sonores en calculant et combinant les valeurs moyennes des axes horizontaux et verticaux [12]. Plus récemment, Woon Seung Yeo et Jonathan Berger proposèrent un système, *SonART*, pour la sonification d'images statiques [21]. Ils introduisirent les concepts de sondage (*probing*) et balayage (*scanning*) qui rapprochent leur travail de la synthèse *wave terrain*. Norihisa Nagano et Kazuhiro Jo, de leur part, ré-interprètent simplement et directement chaque pixel de l'image comme un échantillon sonore dans leur logiciel, et pièce du même nom, *Monalisa* [8].

Il existe de nombreux autres exemples d'applications musicales de techniques d'analyse d'image, mais souvent, les systèmes proposés ne permettent pas l'utilisation d'images arbitraires. Il peut s'agir, par exemple, de systèmes développés spécialement pour la danse, comme le système *EyeCon* de la troupe Palindrome [20], ou bien d'interfaces dédiées comme le *Mouthesizer* de Michael Lyons et Nobuji Tetsutani [13].

2.2. Pourquoi la sonification ?

On peut identifier un certain nombre de raisons pour utiliser la sonification comme outil de composition musicale. Tsékhanovski rêvait à une musique mystérieuse et latente, dissimulée dans les formes plastiques. L'idée que des sons ou des musiques cachées dans les formes du monde pourraient être découverts à l'aide de la technologie apparaît pour la première fois dans les écrits du poète Rainer Maria von Rilke. Dans un court texte de 1919, *Bruit premier* (*Ur-Geräusch*) [16], Rilke décrit comment, d'abord frappé par le caractère plastique du sillon phonographique, il remarqua ensuite sa similitude avec les sutures crâniennes, ce qui le poussa à rêver au « bruit premier » qui pourrait s'y cacher. Chez Tone ou Watts, influencés par le mouvement Fluxus, cette recherche de structures musicales dans les formes de la nature s'inscrit dans une esthétique Cagienne d'abandon de l'égo par l'artiste. La sonification leur permet d'accorder aux caractères chinois ou aux nuages un rôle non seulement thématique, mais aussi créateur.

Pour Kurenniemi et surtout Rokeby, l'analyse d'image se veut d'abord et avant tout un instrument d'interactivité, le geste humain ayant très souvent joué un rôle central dans l'utilisation du *DIMI-O* ou du *VNS*. Cependant, la simplicité relative des deux systèmes ne fixe que très peu de limites dans leurs modes d'opération, permettant autant une interaction directe, presque instrumentale, ou indirecte comme dans *Measure* (1992) de Rokeby [18].

3. L'ANALYSE D'IMAGE

3.1. Réduction

Le problème de la sonification d'images animées peut-être appréhendé en tant que processus de réduction d'information. Une séquence d'image vidéo d'une résolution de 640 par 480 pixels à 30 images par seconde est composée de 9 216 000 éléments (pixels), soit beaucoup plus que les 88 200 éléments par secondes d'un signal audio stéréo échantillonné à 44 100 Hz. Toute tentative de traduction vers le domaine sonore nécessite donc l'élimination d'une partie de l'information visuelle.

Dans sa forme brute, la séquence video est très souvent composée d'une grande partie d'information redondante, par exemple, des régions uniformes (redondance spatiale) ou immobiles (redondance temporelle). C'est sur cette redondance que repose les fondements des techniques de compression vidéo [6] et d'un point de vue technologique la sonification d'images animées proposée dans cet article se rapproche de certains de ces algorithmes.

L'information contenue dans une séquence d'images brute n'est pas seulement redondante, elle est aussi, d'un point de vue sémantique, de bas niveau. Une image numérisée n'est qu'une collection de mesures locales d'intensité. Une être humain pourra y voir un visage, un paysage ou un texte écrit ; c'est donc qu'il est possible d'en extraire des informations d'ordre beaucoup plus élevé. Cependant, il est important de ne pas se laisser tromper par la facilité avec laquelle le cerveau humain acquiert ces informations : elles sont le produit d'une analyse d'une complexité qui dépasse encore de beaucoup les capacités de l'informatique actuelle. Néanmoins, il existe des algorithmes de vision industrielle qui nous permettent de découvrir qu'il y a un visage dans l'image, que ce visage appartient à une certaine personne et que cette personne semble être triste. (Dans le domaine musical et artistique, les travaux de recherche réalisés à *INFOMUS* portant sur l'extraction de paramètres affectifs de la danse sont de cet ordre [2].) Cependant, il s'agit ici d'algorithmes spécialistes. Un programme conçu pour détecter la présence de visages humains dans une image sera incapable de nous informer sur le contenu d'une image représentant un paysage ou une colonie microbienne. Ainsi, comme plus l'analyse est de haut niveau, plus elle est spécialiste, une approche généraliste, pouvant analyser

des images provenant de sources diverses et arbitraires se doit d'être de relativement bas niveau.

3.2. Points d'intérêt

La technique employée pour analyser l'image repose sur la détection de points d'intérêt, (*features*). Bien qu'il existe de nombreux algorithmes de détection de points d'intérêt, de façon générale, il s'agit d'un point possédant un voisinage de faible redondance spatiale. Le terme coin (*corner*) est aussi employé pour désigner ces points, et le plus souvent ils correspondent à des croisements de lignes ou des changements soudains de direction de contour. Un petit nombre de points d'intérêts, entre quelques dizaines ou centaines au plus, peuvent nous informer sur la structure générale d'une image. L'identification de ces points est d'ailleurs le premier stage de certaines analyses d'image de haut niveau, comme la reconnaissance de formes.

Figure 1. Les cercles marqent les points d'intérêt dans cette image.

3.3. Vecteurs de mouvement

Une fois les points d'intérêt identifiés pour deux images consécutives d'une séquence, il est possible d'établir des correspondances d'une image à l'autre. Pour chaque point d'intérêt dans une image, un algorithme cherche à trouver le point d'intérêt dans l'image suivante lui ressemblant le plus. Il existe à cette fin différentes approches [19] mais toutes produisent un ensemble de vecteurs décrivant le mouvement des points d'intérêts d'une image à l'autre. Les vecteurs sont exprimés sous la forme (x, y, α, θ), ou x et y représentent le point de départ du vecteur, α, son amplitude (vélocité) et θ, son angle. Tout comme l'ensemble des points d'intérêt peut nous informer sur la structure spatiale d'une image, le champ de vecteurs de mouvement peut décrire la structure spatio-temporelle d'une séquence d'images.

À ce stade, il serait possible de pousser l'analyse encore plus loin. Il existe, par exemple, des algorithmes permettant la reconstruction de formes tridimensionnelles à partir d'une séquence de vecteurs de mouvement. Il serait aussi possible de chercher à segmenter l'ensemble de vecteurs en grappes correspondant à des objets réels. Cependant, comme il sera démontré, une telle analyse n'est pas nécessaire à nos fins.

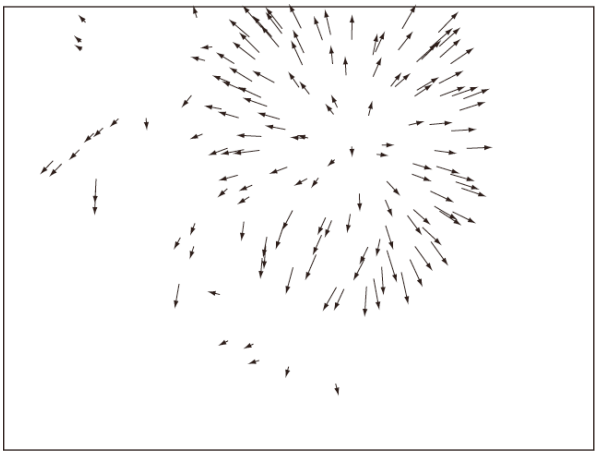

Figure 2. Les flèches représentent les vecteurs de mouvement. Même avec si peu d'information, on peut deviner des feux d'artifice.

4. SONIFICATION

4.1. Vecteurs et synthèse

Une fois les vecteurs de mouvement identifiés, la séquence d'image peut être sonifiée en interprétant chacun de ces vecteurs comme paramètres de contrôle d'un élément sonore d'une synthèse de type microsonique [17], par exemple un partiel dans le cas de synthèse additive ou un grain de synthèse granulaire. Comme le nombre de vecteurs de mouvement dépend surtout de la complexité de la forme visuelle de l'image, on peut déjà affirmer une équivalence : plus la structure spatiale d'une image sera dense, plus le nombre de composantes acoustiques sera grand et donc plus dense aussi sera le son. La quantité, relativement grande, des vecteurs, résout le problème du contrôle d'un grand nombre d'événements qui se pose dans la synthèse granulaire ou additive.

4.2. Association de paramètres

L'association de paramètres spatiaux (position, angle, grandeur) à des qualités acoustiques est réalisée de façon à préserver une similitude morphologique entre l'image et le son tout en laissant un certain degré de liberté au compositeur.

4.2.1. L'espace

La relation la plus évidente est celle de l'espace, car c'est la seule qualité qui soit réellement partagée par le sonore et le visuel. Il convient donc d'associer la position du vecteur de mouvement à celui de l'élément sonore. Par exemple, dans un système stéréo, on utilisera l'axe horizontal de façon à ce qu'un mouvement d'un côté de l'image sera associé à un son positionné au même endroit. Dans un système *surround*, on utilisera autant l'axe horizontal que vertical. Étant donné que le vecteur de mouvement nous informe aussi sur la direction et la vitesse du mouvement, il sera aussi possible de simuler l'effet Doppler. De cette façon, il est possible de générer facilement des trajectoires spatiales qui sont à la fois très complexes et cohérentes.

4.2.2. L'amplitude

Le contrôle d'amplitude s'effectue le plus souvent par la vélocité du vecteur de mouvement. Bien que dans le monde naturel, un mouvement rapide n'est pas nécessairement associé à un grand bruit, c'est une relation qui n'est pas non plus surprenante. En terminologie MIDI le paramètre de vélocité contrôle normalement l'amplitude, et un geste d'archet plus rapide, ou un coup de baguette plus vigoureux est d'habitude accompagné d'un son plus fort. De plus, l'approche présentée ici en est une de sonification du mouvement, donc il se doit qu'une absence de mouvement soit accompagnée par un silence.

4.2.3. Associations arbitraires

Le contrôle de la position et de l'amplitude ne sont pas donc particulièrement problématiques car il est facile d'établir des parallèles entre le sonore et le visuel. Cependant, des paramètres tels que la hauteur, ou – dans le cas de la synthèse granulaire – la transposition, n'ont aucun pendant visuel. Le compositeur se doit donc d'établir des correspondances tout à fait arbitraires pour contrôler ces paramètres. Ainsi, on pourra associer autant la position du vecteur que son angle à la fréquence d'un partiel.

L'aspect complètement arbitraire de certaines associations de paramètres n'est pas une faiblesse. Au contraire, il offre au compositeur une certaine liberté qui peut être jugée nécessaire. Une image, en soi, ne possède pas de son et l'acte de sonification est donc forcément lui-même arbitraire. Il revient donc au compositeur de choisir non seulement quelle image sonifier mais aussi de décider de la meilleure façon de le faire.

5. FORMES

Dans l'approche décrite si-dessus, l'image est réduite à un ensemble de points. Le tout est donc réduit à un ensemble de détails. Dans le cas de la synthèse sonore, un ensemble de détails (partiels, grains) se joint pour former un tout. Si autant l'image de départ et le son produit forment des objets cohérents, leur nature ne peut pas être réduite à celle de leurs éléments. Ce constat est un des fondements de la psychologie de la forme (*Gestalt*) [10].

Comme il a été déjà mentionné, une analyse plus poussée de l'ensemble de vecteurs de mouvements, tout particulièrement la segmentation des vecteurs en grappes, n'est pas nécessaire. Quelqu'un visionnant une séquence d'images pourra très facilement identifier les objets qui s'y trouve. Cette identification est le résultat d'une analyse de haut niveau dont certaines règles sont décrites par les lois de la *Gestalt*. Par exemple si un danseur court d'un côté à l'autre de l'écran, pendant qu'un autre se dirige en direction inverse, nous reconnaîtrons deux formes car leurs détails seront groupés proches l'un de l'autre (loi de la proximité) et se dirigeront dans la même direction (loi de destin commun). Même représentés sous une forme aussi réduite qu'un ensemble d'une centaine de points d'intérêts l'appartenance d'un point à la forme d'un danseur ou de l'autre sera à la fois évidente et entièrement d'ordre perceptuel.

Si on sonifie la séquence d'image représentant ces deux danseurs par synthèse additive, de la manière décrite si-dessus, en associant position avec hauteur, nous obtiendront un certain nombre de partiels positionnés près l'un de l'autre et de hauteur semblable (loi de proximité) et effectuant un glissando et un mouvement spatial dans la même direction (loi de destin commun). On entendra donc deux objets musicaux distincts, effectuant des glissandos et se déplaçant en directions opposées. Ces objets sont eux aussi entièrement perceptuels.

Les lois de proximité et de destin commun peuvent aussi fonctionner de façon inter-modale. On associera donc un objet sonore avec un objet qui lui est proche, ou un geste musical avec un mouvement visuel ayant le même profil dynamique.

6. IMPLÉMENTATION

Il existe présentement deux implémentations du système décrit dans cet article. La première consiste en un *external* pour Max/MSP/Jitter qui calcule les vecteurs de mouvements et qui est accompagné de *patch* démontrant des méthodes de synthèse sonore appropriées. La deuxième, et plus récente, prend la forme d'un logiciel autonome qui analyse des fichiers vidéos et communique, par le protocole UDP, à un environnement de synthèse sonore comme Max/MSP ou Supercollider les paramètres de contrôle générés. Dans les deux cas, l'analyse d'image peut être effectuée en temps réel à plus de 30 images par secondes.

7. DIRECTIONS FUTURES

Le type de point d'intérêt choisi pour le système décrit a l'avantage d'être polyvalent et facile à calculer. Ces points d'intérêts correspondent aussi bien à certains éléments visuellement saillants, ce qui est absolument nécessaire si on désire créer des parallèles perceptifs entre l'image et le son. Cependant, les *corners* ne sont pas le seul type d'élément de bas niveau qui puisse être détecté dans une image. Lignes, taches, textures ou couleurs sont autant d'éléments visuels qui peuvent servir de base à la sonification. Il faudra donc déterminer quel type de synthèse sonore se prête le mieux à ces éléments et quelles associations de paramètres offriront les meilleures possibilités d'expression musicale.

8. RÉFÉRENCES

[1] Avraamov, A. "Синтетическая музыка", *Kinovedcheskie Zapiski*, no. 53, pp.325-333, 2001.

[2] Camurri, A. Hashimoto, S. Ricchetti, M. Ricci, A. Suzuki, K. Trocca, R. and Volpe, G. "EyesWeb: Toward Gesture and Affect Recognition dans Interactive Dance and Music Systems," *Computer Music Journal*, vol. 24, no. 1, pp. 57-69, Apr. 2000.

[3] Chion, M. *L'audiovision : son et image au cinéma.* Nathan Université, Paris, 1994.

[4] Cummings, A. Yasunao Tone. [en ligne] Disponible: www.bbc.co.uk/radio3/cutandsplice/tone2.shtml [Accédé: 29 jan. 2009]

[5] Hacquart, G. *La musique et le cinéma*. P.U.F., Paris, 1959.

[6] Haskell, B.G. Puri, A. Netravali, A. N. *Digital Video: an Introduction to MPEG-2*. Kluwer Academic Publishers, New York, 1996.

[7] Iankovski, B. "Акустический синтез музыкальных красок", *Kinovedcheskie Zapiski*, no. 53, pp.353-367, 2001.

[8] Jo, K. Nagano, N. "Monalisa: See the Sound, Hear the Image", dans *Proceedings of the 2008 International Conference on New Interfaces For Musical Expression*, pp.315-318, 2008.

[9] Kaplan, L. *Laszlo Moholy-Nagy: Biographical Writings*. Duke University Press, North Carolina, 1995.

[10] Koffka, K. *Principles of Gestalt Psychology*. Harcourt, New York, 1967.

[11] Kramer, G., Walker, B., Bonebright, T., Cook, P., Flowers, J., Miner, N.; Neuhoff, J., Bargar, R., Barrass, S., Berger, J., Evreinov, G., Fitch, W., Gröhn, M., Handel, S., Kaper, H., Levkowitz, H., Lodha, S., Shinn-Cunningham, B., Simoni, M., Tipei, S. "The Sonification Report: Status of the Field and Research Agenda", Report prepared for the National Science Foundation by members of the International Community for Auditory Display. ICAD, Santa Fe, NM, 1999.

[12] LaBelle, B. Background Noise, Continuum International Publishing Group, London, 2006.

[13] Lyons, M. J. Tetsutani, N. "Facing the Music: A Facial Action Controlled Musical Interface", dans *Proceedings, CHI 2001, Conference on Human Factors in Computing Systems*, pp. 309-310, 2001

[14] M. Ojanen, J. Suominen, T. Kallio, and K. Lassfolk, "Design principles and user interfaces of Erkki Kurenniemi's electronic musical instruments of the 1960's and 1970's," dans *Proceedings of the 2007 International Conference on New Interfaces For Musical Expression*, pp. 88-93, 2007.

[15] Pelletier, J.M. "Sonified Motion Flow Fields as a Means of Musical Expression", dans *Proceedings of the 2008 International Conference on New Interfaces For Musical Expression*, pp. 158-163, 2008.

[16] Rilke, R. M. *Oeuves en prose*. Gallimard, Bibliothèque de la Pléiade, Paris, 1993.

[17] Roads, C. *Microsound*. MIT Press, Cambridge Mass. 2001.

[18] Rokeby D. Installations. [en ligne] Disponible: homepage.mac.com/davidrokeby/installations.html [Accédé: Jan. 29 2009]

[19] Smith, P. Sinclair, D. Cipolla, R. and Wood, K. "Effective Corner Matching," dans *Proceedings of the 9th BritishMachine Vision Conference*, pp. 545–556, 1998.

[20] Watts, R. Behrman, D. Diamond, R. Cloud Music. [en ligne] Disponible: www.vasulka.org/Kitchen/PDF_Eigenwelt/pdf/152-154.pdf [Accédé: Jan. 29 2009]

[21] Wechsler, R. Weiß, F. Dowling, P. "EyeCon – A Motion Sensing Tool for Creating Interactive Dance, Music and Video Projections" dans *Proceedings of the Society for the Study of Artificial Intelligence and the Simulation of Behavior (SSAISB)'s convention: Motion, Emotion and Cognition at University of Leeds, England, March 29, 2004,* 2004

[22] Yeo, W. Berger, J. "Application of Image Sonification Methods to Music", dans *Proceedings of the 2005 International Computer Music Conference*, 2005.