SYNTHÈSE CONCATÉNATIVE ET IMAGE : DYCI2 ET OMAX-VIDÉO

Georges Bloch
GREAM
IRCAM
gbloch@unistra.fr

Jérôme Nika
IRCAM
Jerome.Nika@ircam.fr

Quentin Barrois
UR 3402 – ACCRA
barroisquentin@gmail.com

RÉSUMÉ

Ce texte étudie la relation entre logiciels de la famille *OMax-DYCI2* et l'image (*OMax_Vidéo*) au cours de leur évolution.

Ces outils permettent de développer divers processus d'images asservis aux sons. Le statut de l'image varie : il va d'une représentation de l'interprète à un plaquage forcé de l'image sur la musique. Parmi les choses possibles, on peut rendre visible une construction musicale ; « écouter » le son d'un film pour déclencher une autre musique. On peut aussi écouter une musique afin d'assembler les extraits d'un film dont la bande-son aura des points communs avec la musique que l'on écoute. Enfin, on peut enchaîner ou mettre en boucle ces divers processus.

Le logiciel permet aussi d'utiliser un *scénario*, plan à plus ou moins long terme qui pilote l'improvisation à partir de la mémoire sonore. Le scénario utilise une notation symbolique formée du même alphabet que celui de l'analyse qui organise la mémoire sonore. L'alphabet peut être généré de façon automatique à l'aide de descripteurs spectraux.

Enfin, on peut envisager d'appliquer directement aux images ce système d'analyse et de scénario.

Des projets comme *Paris bout-à-bout* ou le projet *Ozu* illustrent ces péripéties sonores et visuelles.

1. INTRODUCTION

Les logiciels de la famille *OMax-DYCI2* existent depuis près de vingt ans, leur extension vidéo depuis une quinzaine d'années. Ces logiciels permettent à une machine d'improviser sur un répertoire ou une improvisation acquise à la volée, en faisant des propositions qui peuvent être à la fois inattendues et stylistiquement cohérentes.

Ces logiciels pratiquent la synthèse concaténative : la musique produite est avant tout une *recombinaison* de la musique mémorisée et annotée. De ce fait, il est facile de prévoir une extension vidéo, pour peu qu'il existe – ou que l'on définisse – une image synchrone à la bande son. Ce système encourage même un type de création assez peu commun : une image pilotée par la musique sans qu'elle en soit nécessairement une illustration.

Comme nous le verrons, les évolutions du système – entre autres, l'utilisation d'un scénario et le statut variable de la mémoire sonore – modifient les rapports possibles entre image et son. Enfin, puisque l'on parle de synthèse concaténative, donc de montage, et de scénario temporel, nous pouvons envisager d'élargir le système à des paramètres purement visuels.

2. OMAX/DYCI2 : SYNTHÈSE CONCATÉNATIVE

La famille *OMax* de logiciels d'improvisation assistée par ordinateur existe depuis le début du siècle et a connu de nombreux développements [1]. Tout est parti d'une recherche sur la modélisation stylistique : un modèle stylistique de musique implémenté dans une machine peut être jugé valide si la machine est capable de ressortir une musique originale dont le style est cohérent avec celui de la musique qui lui a été apprise. L'enthousiasme immédiat que ces « machines infernales [2] » ont provoqué chez certains jazzmen – il faudrait au moins citer Bernard Lubat et Philippe Leclerc – a rapidement fait évoluer le projet. Par la suite, la recherche s'est donc beaucoup focalisée sur la relation entre l'interprète et la machine [3].

2.1. Concaténation

Même si certaines des premières versions d'*OMax* comportaient des modules de génération (mettant notamment en œuvre les accords de substitution en jazz), *tous* les logiciels de la famille improvisent avant tout selon un modèle de synthèse dite « concaténative » : la musique générée est une recombinaison de ce que l'on appelle la « mémoire », en l'occurrence une musique annotée qui a été entrée dans la machine avant le concert, ou, le plus souvent, au cours du concert. On recolle donc des bouts de musique existants. Si rien n'a été stocké en mémoire avant le début du concert, la machine ne produit rien au début car sa mémoire est vide ; ensuite, elle joue exclusivement à partir du matériau que l'interprète a joué auparavant sur scène.

Cette mémoire est toujours décrite par une représentation *symbolique* de la musique jouée. Cette représentation peut prendre des formes diverses : notes, descripteurs spectraux ou toute combinaison multidimension-

[1] On en voit les prémisses dans [16]. Pour les développements parallèles, voir par exemple *Py-Oracle* [22] et *Mimi4x* [17].

[2] L'expression est de Hervé Sellin.

[3] Voir les travaux de Marc Chemillier, par exemple dans [2].

nelle de ces paramètres. Le caractère symbolique, notons-le, est inhérent aux processus de concaténation : un procédé aussi simple qu'un *looper* décrit une mémoire à partir de deux symboles, le temps de début et le temps de fin de la boucle. Notons que, au cinéma, un plan de montage est aussi une représentation symbolique… décrivant la synthèse concaténative qu'est le montage.

Pour la famille *OMax-DYCI2*, le modèle de mémoire repose donc sur une ou plusieurs segmentations liées à la valeur de certains paramètres (par exemple des notes, des temps, des MFCC, des pics de *loudness*), ainsi qu'une cartographie de ces segments avec leurs valeurs. La catégorisation des valeurs – la création d'un alphabet – est déjà une cartographie ; dans le cas d'*OMax*, celle-ci est nettement améliorée par l'utilisation de l'oracle des facteurs, lequel repère et relie les répétitions à la volée [1].

L'architecture générale du système n'est pas si lointaine du schéma de musique interactive décrit par Chadabe dans [9]. Ce qui en diffère est lié au modèle de mémoire (en gris sur la figure) imposé par la logique de concaténation. On notera que les interprètes (en rouge) produisent la musique et peuvent aussi piloter la stratégie d'improvisation [4]. On peut également rentrer du répertoire à l'avance dans la mémoire. Enfin, l'exécution peut être asservie à un texte (partition).

Ce schéma ne représente pas l'interaction la plus banale, sans doute la plus importante : l'écoute par les musiciens du son qui sort de la machine (flèche bleue en bas à droite). La machine peut éventuellement proposer aux interprètes un retour symbolique sous forme d'indications (jaune).

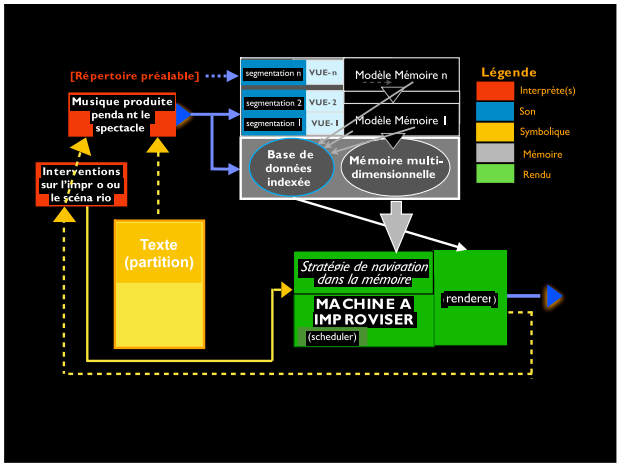

Figure 1. Ce schéma de base d'OMax pourrait s'appliquer à tout processus interactif utilisant la synthèse concaténative.

2.2. Scénario et rupture

Mais nous voulons aussi envisager un volet compositionnel, afin de rechercher ce mouton à cinq pattes que l'on appelle l'improvisation composée : on s'efforce de profiter de la puissance abstraite d'une structure à long terme tout en conservant l'expressivité instantanée de l'improvisation [5]. La famille *OMax-DYCI2* a rapidement reconnu la nécessité d'un scénario pour piloter la navigation dans la mémoire.

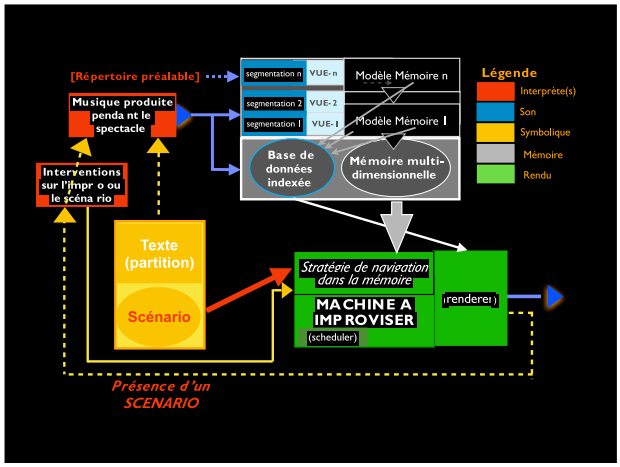

Figure 2. L'ajout d'un scénario imposant une structure sur le long ou moyen terme. Il pilote la navigation dans la mémoire.

L'exemple de la grille de jazz, texte qui impose une structure sur une improvisation, est lumineux. Mais il est aussi juste que trompeur : juste, parce ces grilles ont été parmi les premiers scénarios utilisés et ont suscité d'importants développements – notamment la thèse de Jérôme Nika [19] ; faux, parce que, comme l'explique fort bien Gilles Deleuze [6], c'est par la rupture que la répétition permet le dépassement de la mémoire. La plupart des processus compositionnels utilisent la répétition pour construire un point de rupture.

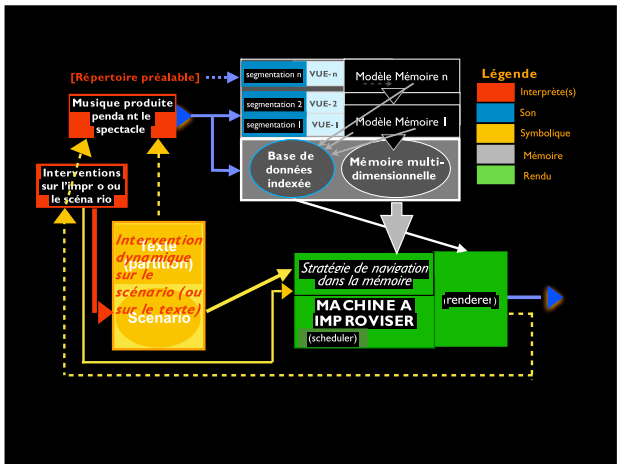

Figure 3. Scénario dynamique modifiable à la volée.

Mais rien n'oblige le scénario à éviter les points de rupture. Ces derniers sont d'autant plus faciles à créer si l'on a la possibilité de modifier le scénario de façon dynamique au cours du spectacle.

Entre autres choses, *DYCI2* permet de modifier le scénario de façon dynamique, et de relier directement la

[4] Cette stratégie de navigation est décrite en [3]. Ce schéma sous-entend des interprètes instrumentistes et d'autres derrière l'ordinateur.

[5] Cette problématique est déjà bien présente dans [5].

[6] Dans le chapitre 2 « La répétition pour elle-même » de [14]. C'est ce que Deleuze appelle la troisième synthèse du temps.

fabrication de la mémoire et celle du scénario. Ces propriétés seront décrites dans la partie 4.

3. UNE IMAGE GUIDÉE PAR LE SON ?

3.1. L'évidence de la vidéo

À partir de 2004, lorsque *OMax* a fonctionné avec de l'audio et non plus seulement avec du MIDI, la présence d'une extension vidéo tombait sous le sens.

Il s'agit de concaténer les éléments d'une mémoire : si images il y a, il s'agit donc de les monter [6]. Par conséquent, si l'on enregistre l'image en même temps que le son et que l'on dresse un tableau de correspondance de *time-codes* entre l'image et le son, à toute date de son dans la mémoire correspondra une image.

La première idée qui nous est venue à l'esprit a été de filmer l'interprète en action. L'improvisation mise en œuvre par la machine devient alors *vue* autant *qu'entendue*, avec toute la magie de l'infaisable qu'elle est susceptible de véhiculer (transitions, doigtés, etc.). Le premier prototype a été monté avec la collaboration d'étudiants de l'université de Strasbourg et a fait l'objet d'une présentation dès 2005 [6].

Figure 3. Philippe Leclerc jouant avec lui-même : *OMax_vidéo* en 2006. On voit à l'écran une improvisation de la machine à partir de ce qu'il a joué auparavant.

Un autre intérêt de la vidéo – peut-être d'ordre plus scientifique – était de rendre visibles des points de montage qui, parfois, n'étaient pas ou peu audibles.

Mais la possibilité d'improviser également à partir de fichiers audio existants permettait de faire appel à des vidéos du répertoire, et donc de générer de nouvelles improvisations de légendes comme Thelonius Monk, Jimmy Hendrix ou Ella Fitzgerald ou, mieux encore, de jouer en duo avec ces géants : cela reste l'utilisation la plus fréquente du logiciel avec l'extension vidéo.

Mais qu'en est-il d'images ne représentant pas les interprètes ? Qu'en est-il, par exemple, de musiques de films ? Les premières expériences ont été plutôt d'ordre expérimental. Elles ont permis de tester quelques hypothèses sur les relations entre images et sons :

— Faire tourner la machine sur une musique très redondante à l'image devient vite assez ridicule. Cette expérience faite, par exemple, sur un *teaser* de *Metropolis* montrait que le leitmotiv « carte de visite » présentait le même défaut à l'envers (dans le sens sonore–visuel) : une répétition musicale relativement inattendue devient banale lorsqu'elle est associée au retour automatique d'une même image, laquelle lui colle alors de façon pléonastique ;
— Faire tourner une improvisation à partir de musiques complètement étrangères à l'image démontre le principe des points de synchronisation accidentels ;
— Ce dernier montre aussi la faiblesse des théories topiques [18] appliquées à la musique de film, puisqu'une association immédiate entre musique et image s'effectue par simple répétition, même en l'absence d'un quelconque marqueur topique.

On résume en table 1 les diverses relations entre la musique montée par le processus et les images dont le montage est asservi au montage musical.

Statut de l'image	Lien : mémoire acquise en concert	Lien : mémoire pré-chargée (répertoire)		Sans lien
Exemples	On filme l'interprète qui improvise	Interprètes (film de concert)	Film (musique)	Temps et motifs musicaux imposés

Table 1. Les divers types d'images asservies au montage musical.

Mais ces expériences ont rapidement pris une autre tournure, avec l'apparition de scénarios, puis d'outils de guidage associés ou non à des descripteurs spectraux.

3.2. Dispositif vidéo actuel

La version actuelle d'*OMax_Vidéo* est une bibliothèque *Max–Jitter* permettant :

— l'acquisition de données vidéo en synchronisation avec *OMax* ou *DYCI2* (avec création d'une table de correspondance) ;
— la lecture de films dont le montage est synchronisé avec le montage son : soit celui acquis durant le concert, soit des films correspondant à des mémoires présentes auparavant ;
— dans le cas d'un ciné-concert, la création d'une table de synchronisation entre le son et l'image. Cette table permet, lorsqu'on lance une improvisation sonore, de remonter les images du film initialement associées aux sons. Le montage visuel est alors asservi à la logique sonore.

Le dispositif actuel permet d'avoir plusieurs sources synchronisées à des voix *OMax* ou *DYCI2* ou des sources dites « Triches » non synchronisées (voir figure 4). On peut distribuer ces quatre sources sur 1, 4, 9 ou 16 écrans et leur appliquer divers effets vidéo. On peut aussi capter une source en temps réel, synchronisée à ce qui est joué.

Les pistes de type « Triche » sont a priori utilisées pour passer un film indépendant de la captation musicale. Le mot a été employé dans les premières versions d'*OMax* afin de mettre en valeur la synchronisation.

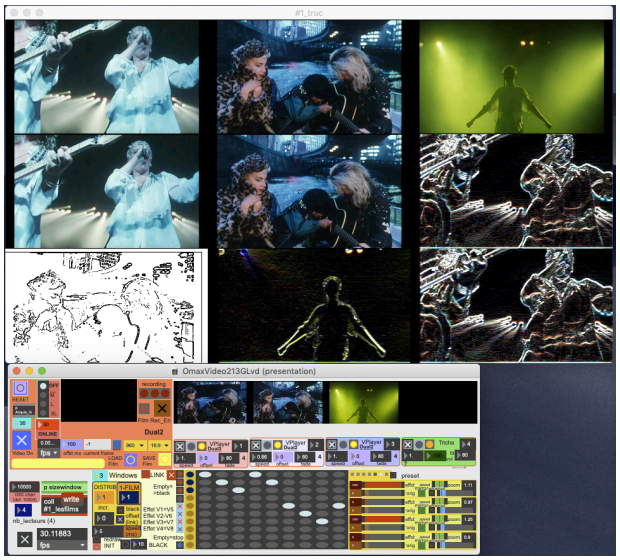

Figure 4. Capture d'écran d'*OMax_Vidéo*. La piste verte à droite de la fenêtre inférieure est une piste « Triche », non synchronisée.

Néanmoins, on peut aussi utiliser ces pistes en situation de ciné-concert – lorsque la musique est produite devant des images préexistantes – afin de *synchroniser* dans la mémoire les passages des films joués avec la musique interprétée en direct. Si l'on utilise ensuite la machine pour faire un montage sonore, on peut projeter les images d'origine correspondantes. Bien que toujours asservi à la musique, ce montage ajoute à l'image un statut supplémentaire : celui où la mémoire acquise au concert provient d'une exécution *live* devant le film. L'image est toujours liée à la mémoire acquise en concert, mais elle préexiste à cette mémoire. La table 2 intègre cette possibilité à celles de la table 1.

Lien : mémoire acquise en concert		Lien : mémoire pré-chargée (répertoire)		Sans lien
On filme l'interprète qui improvise	Ciné-concert (images pré-existantes)	Interprètes (film de concert)	Film (musique)	Temps et motifs musicaux imposés

Table 2. Types d'image asservie au montage musical.

3.3. Images asservies à la musique

Le projet *Three Ladies* a été conçu en 2015 pour le festival *Le pietre che cantano* à L'Aquila (Italie), dans le cadre du centenaire de Billie Holiday, Edith Piaf et Elisabeth Schwarzkopf. L'idée était de faire chanter ces trois dames ensemble.

La chose a été possible en associant des interprétations de Piaf et de Schwarzkopf à un scénario fondé sur des standards de jazz. Le caractère étonnant du mélange et la réaction en direct du pianiste Hervé Sellin se traduit par une sonorité peu usuelle [7] (figure 5).

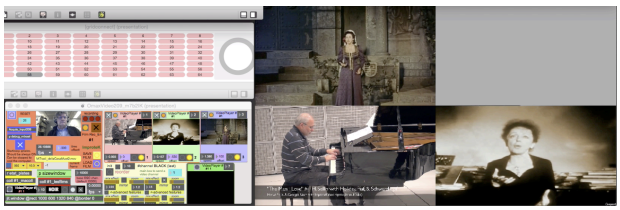

Figure 5. Le scénario est *The Man I Love*. On en est à la deuxième mesure de la dernière phrase de la grille. Billie Holiday chante, mais on ne la voit pas. On voit par contre Piaf, Lisa della Casa et Hervé Sellin qui suivent la même grille harmonique. La « mémoire » de Della Casa est « mi tradi » de *Don Giovanni*, celle de Piaf *Mon Dieu*.

Disons-le : il n'y avait pas de vidéo dans le projet d'origine ; la vidéo a été conçue ensuite comme une application des nouvelles architectures informatiques, celles, musicales, mises en œuvre par Jérôme Nika [19] et celles pour l'image utilisant OpenGL [7]. Ce type de procédé a été largement employé depuis : par exemple dans *La Meute Kitsch* (avec Hervé Sellin) au colloque Abraham Moles à l'université de Strasbourg en 2016, ou *Til Midnight* (avec Rémy Fox) au colloque Improtech à Philadelphie en 2017. Ces spectacles utilisent des images liées à des mémoires préchargées. Après 2016, le mélange entre films musicaux (avec les images des interprètes) et musiques de films (avec l'image du film) est fréquent : *La Meute Kitsch* mêle des personnages qui jouent ou qui chantent – de John Coltrane au petit chaperon rouge de Tex Avery – et des musiques de films de fiction (Leone–Morricone) [8].

Le dispositif *Three Ladies* qui a été présenté à Paris, au festival Manifeste, le 5 septembre 2020, a fait appel, entre autres, à un dispositif semblable.

Enfin, dans tous ces exemples, l'image est asservie au son, le plus souvent au nom d'un scénario musical.

4. *DYCI2* ET APRÈS

DYCI2 est la bibliothèque la plus utilisée dans le cadre d'une utilisation avec la vidéo. DYCI2 est une bibliothèque d'agents génératifs pour la performance et l'interaction musicale combinant les approches libres, planifiées et réactives de la génération à partir d'un corpus, ainsi que des modèles de scénarios dynamiques à court terme (« meta-Djing »). DYCI2 consiste en :

— une bibliothèque Python définissant des modèles et des outils pour la génération de séquences créatives à partir de modèles de séquences. Elle implémente plusieurs modèles, heuristiques, stratégies de gestion du temps, et architectures d'agents interactifs ;
— une bibliothèque Max d'agents temps-réel interfacée avec la bibliothèque Python [9].

[7] Extrait sur `www.youtube.com/watch?v=UUXDkdt76J8`.
[8] Visible sur `vimeo.com/237664490`.
[9] Voir `github.com/DYCI2/Dyci2Lib`.

4.1. Stratégies de génération musicale

À partir d'un agent embarquant une « mémoire » musicale, *DYCI2* propose différentes stratégies génératives.

4.1.1. Génération libre

Une fois le modèle de mémoire construit, une requête dite « free » entraîne la génération d'une séquence musicale inédite selon la logique temporelle interne du matériau appris, sans autre contrainte de structure ou d'écoute réactive. Cette stratégie est semblable à celle de OMax, légèrement simplifiée : la totalité des stratégies de navigation décrites en [3] n'y figure pas (cf. figure 1).

4.1.2. Guidage par scénario long-terme

La génération peut être guidée par un scénario déterminé par l'utilisateur. Cette séquence symbolique de référence est définie sur le même alphabet que les annotations de la mémoire musicale utilisée (labels d'accords, modes de jeu, etc.) et guide l'improvisation (séquence d'accords, séquence de modes de jeu, etc.). Il s'agit donc de trouver des segments de la mémoire correspondant aux portions successives du scénario à suivre et de les enchaîner de manière créative. Pour atteindre cet objectif, à chaque instant de la génération, le modèle proposé associe l'*anticipation* en assurant la continuité avec le futur du scénario, et la *cohérence avec la logique musicale de la mémoire* (cf. figure 2).

4.1.3. Guidage par scénario dynamique à court terme

L'implémentation des stratégies de gestion de requêtes concurrentes et de l'approche de la réaction en tant que réécriture d'anticipations préalablement générées a permis d'introduire la notion de scénarios dynamiques à court terme. Au cours de la performance, un opérateur-musicien peut ainsi envoyer des requêtes correspondant à des séquences de labels spécifiant ce que l'agent doit générer et jouer sur-le-champ, tout en maintenant la cohérence avec ce qui vient d'être joué ainsi qu'avec les anticipations préalablement générées dans le cas où une révision des anticipations est nécessaire. Ce paradigme de guidage introduit un mode de jeu que l'on pourrait qualifier de « meta DJing » ou « DJing d'intentions » : en effet, un opérateur-musicien peut improviser en contrôlant un agent à l'échelle de la narration musicale. Par exemple, on envoie cette instruction : « à partir du temps prochain : générer et jouer une séquence correspondant à la suite d'accords Dm7 G7 CMaj7 » ; ou celle-ci : « maintenant : générer et jouer une séquence partant de *grave rugueux* pour arriver à *aigu brillant* » (cf. figure 3 qui ajoute cette intervention dynamique sur le scénario déjà présent dans la figure 2).

4.2. Utilisation de descripteurs audio

L'utilisation de descripteurs audio permet, entre autres, d'automatiser certaines tâches, ce qui modifie profondément le système.

4.2.1. Guidage par scénario de descripteurs audio

Les stratégies de guidage par scénario ou par scénarios dynamiques à court terme peuvent également être utilisées pour naviguer dans une mémoire constituée par l'analyse automatique d'un fichier ou d'un flux audio. Au cours de la performance, l'utilisateur peut ainsi envoyer des requêtes spécifiant des séquences de classes que l'agent doit générer et jouer. Les classes constituant l'alphabet de labels sont obtenues par une analyse du fichier « mémoire » selon une sélection de descripteurs effectuée par l'utilisateur, puis par un *clustering* discrétisant l'espace de descripteurs en un nombre de classes également choisi par l'utilisateur. On est dans le même cas que précédemment (figure 3), mais la génération du vocabulaire d'analyse est automatique ; répétons-le, c'est ce même vocabulaire qui sera utilisé pour le scénario. Dans ce cas, la mémoire est analysée *à l'avance*. On prend un passage, soit chargé à l'avance dans le répertoire de la machine, soit joué *live* auparavant mais considéré comme clos.

4.2.2. Guidage par écoute réactive

La navigation dans une mémoire ainsi constituée par analyse automatique peut également être pilotée par un module d'écoute/analyse temps-réel. Un flux audio capté en temps réel – par exemple un musicien co-improvisant avec l'agent – est analysé selon la même sélection de descripteurs que la mémoire pour créer automatiquement des requêtes de génération (cf. figure 6).

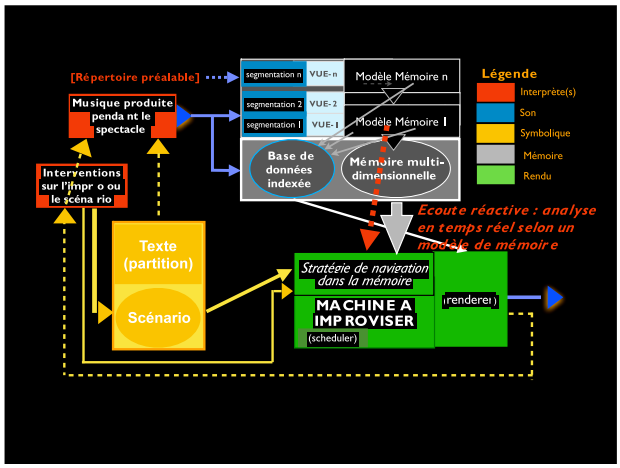

Figure 6. Écoute réactive. Cette écoute obéit aux paramètres de l'analyse *préalable* de la mémoire, laquelle fixe les paramètres avec lesquels le jeu en temps réel de l'interprète sera analysé.

4.2.3. *Guidage hybride écoute réactive/scénarios*

La fusion de cette dernière stratégie avec le paradigme de scénarios à court terme a également permis d'introduire de nouvelles stratégies de génération pilotées par l'écoute, entre lesquelles un agent peut alterner au cours d'une même performance : le label déterminé par l'écoute peut être utilisé pour créer un scénario court terme dans lequel il est répété, dont il est le point de départ, ou encore dont il est la destination. Ces stratégies introduisent un contrôle hybride entre le musicien produisant le stimulus et l'opérateur-musicien « composant » en temps réel les modalités de la réactivité dans une dynamique de compagnonnage humain / humain–machine.

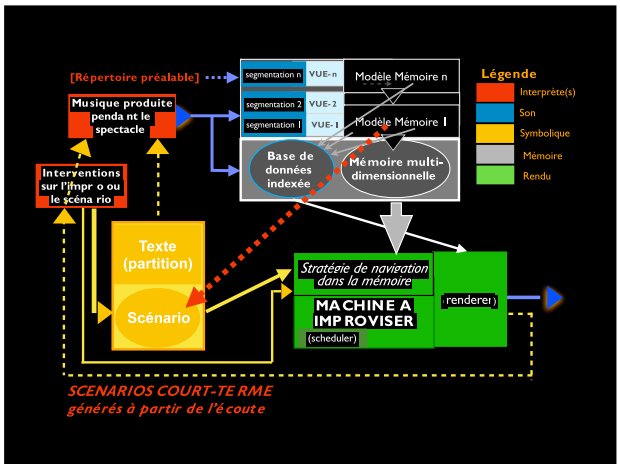

Figure 7. Guidage hybride entre écoute réactive et scénario : on crée un scénario à court terme qui répète des sons de la mémoire correspondant aux labels analysés à l'entrée, se dirige vers eux ou s'en s'éloigne.

4.3. MuBu/CataRT

Un travail mené en collaboration avec Diemo Schwarz et l'équipe ISMM de l'Ircam a conduit au développement du nouveau module d'analyse/rendu utilisant l'environnement *multi-buffer MuBu* [10]. *MuBu* est un conteneur de données sonores et de mouvement fournissant de la mémoire structurée pour le son et le mouvement enregistrés à travers des interfaces et des opérateurs en temps réel en tant qu'objets externes complémentaires pour Max. *CataRT by MuBu* [11], développé au sein de l'environnement *MuBu*, est un système de synthèse concaténative temps-réel qui permet de jouer des « grains » sonores (à partir d'un grand corpus de sons segmentés et analysés par descripteur) selon la proximité avec une cible dans l'espace descripteur, à l'aide d'une souris ou d'un contrôleur externe. Cette approche peut être vue comme une extension de la synthèse granulaire donnant accès à des caractéristiques sonores spécifiques. L'interaction repose sur une interface simple consistant en l'affichage d'une projection 2D de l'espace de descripteurs, et une navigation avec la souris, où les grains sont sélectionnés et joués par proximité géométrique. Cette combinaison des recherches issues du projet *DYCI2* et du moteur *CataRT* a permis la création de nouvelles stratégies de génération : le guidage par scénarios de descripteurs audio et le guidage par écoute réactive. Notons que cette nouvelle technologie enrichit *CataRT* d'une prise en compte de la temporalité du fichier « mémoire » et d'une optimisation des chemins choisis dans l'espace de descripteurs.

4.3.1. *Génération libre structurée*

La modularité de la bibliothèque permet de chaîner deux (ou plusieurs) agents de manière à ce que le type de contenu retourné par le premier soit le type de labels utilisé pour construire les requêtes pilotant le second. Ce type d'utilisation permet d'ajouter une dimension verticale à l'improvisation, avec, par exemple, une chaîne dans laquelle un agent se spécialise dans l'harmonisation et un second dans l'arrangement. Il permet également d'introduire un intermédiaire entre les stratégies de génération « libre » et « structurée ». Un premier agent peut ainsi, par exemple, naviguer de manière « libre » dans sa mémoire et retourner non pas les contenus musicaux des états par lesquels il passe, mais les labels associés. La séquence de labels ainsi construite sera donc inédite, mais cohérente avec la structure de sa mémoire, et pourra servir de scénario pour un autre agent générant du contenu musical.

4.3.2. *Développements actuels et futurs*

La représentation de mémoires multidimensionnelles et les méthodes de navigation dans ce type de données ouvre de nombreuses perspectives [12, 20]. Une autre direction explorée est celle de guider la génération par des scénarios à court terme, non plus seulement déterminés par un utilisateur, mais également inférés d'une écoute en temps réel d'un musicien [9, 10].

5. ÉCOUTER L'IMAGE ?

OMax-Video et *DYCI2* permettent d'imaginer de nouvelles relations entre l'image et le son, où le son est moteur de plusieurs manières. C'est le cas du projet de film interactif *Paris bout-à-bout*.

Mais on peut même envisager d'utiliser le principe de la mémoire et du scénario pour générer des « improvisations » filmiques selon des critères liés à l'image. On peut même envisager à terme que ces critères soient déduits de manière automatique à partir d'une analyse de l'image.

5.1. « Paris bout-à-bout »

Paris bout-à-bout est un ciné-concert interactif. Du fait de l'utilisation d'*OMax* et de *DYCI2*, la musique jouée pendant le concert a un effet sur le déroulement des images.

[10] Voir `forumnet.ircam.fr/fr/produit/mubu`.
[11] Voir `imtr.ircam.fr/imtr/CataRT`.

Les images sont des plans-séquences filmés par la cinéaste Nurith Aviv dans divers lieux de Paris, notamment le métro, ainsi que des photos de Paris plus récentes. La prochaine version prévoit d'y ajouter un plan-séquence filmé par la cinéaste en *live* au cours du concert, qui sera réinjecté en temps réel dans le dispositif du spectacle.

Une version de *Paris bout-à-bout* (sans prise de vue en direct) a été présentée à Athènes, au festival Improtech, le 28 septembre 2019, avec Jaap Blonk, Hervé Sellin et Georges Bloch.

5.1.1. Ciné-concert en montage direct

Les images d'origine filmées par Nurith Aviv proposent une promenade poétique dans le Paris de 1993. Outre plusieurs stations de métro, on peut y voir la boutique d'un fleuriste, une piscine, une arcade de jeux vidéo, un cabinet d'échographie, un salon de coiffure, une galerie, des scènes de rue, etc. Chaque scène est constituée d'un unique plan-séquence. La durée de ces plans varie d'une à cinq minutes.

Le choix de la longueur et de l'ordre des plans fait l'objet d'une première décision, ne serait-ce qu'en fonction de la durée du spectacle. En conservant l'idée du plan-séquence (si on coupe le plan, c'est au début et/ou à la fin), on crée le scénario fondamental du film.

5.1.2. Image guidée par le son : improvisation filmique à partir de la mémoire musicale du ciné-concert

Le film projeté est chaque fois légèrement différent : il y a des passages improvisés, des points d'orgue, etc. Comme expliqué en 3.2, une table de synchronisation est construite entre le son joué et le film projeté. Cela permet de projeter à un moment des images « improvisées » à partir des paramètres musicaux qui ont été acquis lors du ciné-concert. C'est le cas vers la fin du spectacle, qui comporte une assez longue séquence improvisée par la machine ; les interprètes ne jouent pas, puis se joignent à la partie improvisée.

Deux remarques :

– Dans un film composé de plans-séquences, ces parties improvisées sont très évidentes puisque, dans ces parties, la longueur des plans devient tout à coup plus classique – grosso modo, entre une et dix secondes ; de plus, on revoit un montage réalisé à partir d'images déjà vues ;
– Cette partie « improvisée » entre elle-même dans la table de synchronisation. La machine peut ensuite improviser – dans une mise en abyme – à partir de cette improvisation.

5.1.3. Ré-injection d'une image associée auparavant avec un extrait de musique ou de bruitage issu de la machine

On est dans un cas similaire au cas précédent, mais on voit toujours le film du ciné-concert ; l'image ajoutée par la machine modifie la perception de la musique.

Le film commence sur un plan d'échographie, accompagnée par les sons de bouche de Jaap Blonk, plutôt comiques (cf. figure 8).

Figure 8. Au début de *Paris bout-à-bout*, Jaap Blonk (à droite) accompagne de bruits de bouche le plan d'échographie visible à l'image (image Jeff Joly).

Plus tard dans le film, il y a une séquence de jeux vidéo d'arcade (on est en 1993 !). Le son du film et le bruitage « moteuristique » de Jaap Blonk sont renforcés par une improvisation de la machine sur le passage précédent. La mémoire musicale est celle de la scène d'échographie, dans laquelle l'ordinateur navigue librement en cherchant les transitions les plus fluides.

Il y a une très grande cohérence entre ces sons de bruitage, ne serait-ce que parce qu'ils sont produits par la même voix. L'ajout d'une piste sonore, voire de deux, amplifie le brouhaha de l'arcade et souligne la passion de ces jeunes gens accrochés à leurs écrans.

Tout change si l'on montre l'image (d'échographie) qui correspondait à l'origine au son utilisé. Le terme de « correspondance » n'est d'ailleurs pas rigoureusement exact : cette image est celle qui est *à l'origine* de ce son. La montrer provoque évidemment un changement de perception visuelle, même si on la montre en même temps que le film d'origine qui continue de tourner. Surtout, cette image dérange par sa logique sonore par trop évidente, perceptible, une fois encore, par la fréquence des points de montage, à l'opposé d'un film construit sur des plans très longs.

Figure 9. *Paris bout-à-bout* : Jaap Blonk bruite un quad sur une piste dans un jeu vidéo : peu après, l'engin va s'écraser contre un obstacle et exploser. Le bruitage est soutenu par une improvisation de la machine sur les sons de la scène du début, mais… ce montage sonore est *vu* en haut et à gauche de l'écran (image Jeff Joly).

5.1.4. Réaction au son du film ou des musiciens pour déclencher une musique

Le film d'origine est *sonore*. Il est possible de l'écouter, et de générer de la musique.

Entre autres choses, *Paris bout-à-bout* utilise un morceau composé par Hervé Sellin, *Always too Soon*. Jouée au piano solo pendant une séquence, cette pièce revient plusieurs fois, mais sous forme de fragments d'un enregistrement de la pièce en quintette [21]. L'un des plans du film montre un vernissage dans une galerie parisienne. Les bavardages des convives (et les éventuels ajouts sonores de Jaap Blonk) déclenchent des fragments de cette version enregistrée, qui agit à la fois comme un ensemble jouant dans une salle de fond, une réaction aux voix des personnes présentes et une sorte de thème conducteur du ciné-concert.

5.2. Scénario d'images-moteur

La présence de scénarios permet d'utiliser des critères filmiques, à condition que l'on soit capable d'annoter le film. Le projet *Paris bout-à-bout* a aussi utilisé de tels critères. *DYCI2* autorise un scénario comptant jusqu'à cinq critères différents. La table 3 présente une analyse du début du plan dans la station de métro Nation.

Les cinq critères retenus sont :

1. Mouvement de caméra par tranche de 45° ;
2. Déplacement du sujet filmé par tranche de 90° ;
3. Sujet : 1 (1) ou plusieurs (2) personnes, écran (e) ou rien (0) ;
4. Métro : non (0), station (1), métro présent (2) ;
5. Longueur du plan. Le test 0 < 0,75 s < 1,5 s < 3 s définit quatre catégories : 500, 1000, 2000, 4000.

La valeur NIL est utilisée si un critère est inopérant.

				camera 0-45-90=> 360	source 0-90=> 360	sujet 0/1/2/e	long plan	metro 0/1/2	CRITERES
TC			TC (ms.)	CR1	CR2	CR3	CR4	CR5	
4	21	7	261280	NIL	NIL	2	4000	1	261280; NIL NIL 2 1 4000
4	24	12	264480	225	NIL	2	4000	1	264480; 225 NIL 2 1 4000
4	28	5	268200	225	NIL	2	4000	1	268200; 225 NIL 2 1 4000
4	31	20	271800	270	NIL	2	4000	1	271800; 270 NIL 2 1 4000
4	36	1	276040	315	NIL	2	2000	1	276040; 315 NIL 2 1 2000
4	38	14	278560	315	NIL	NIL	2000	1	278560; 315 NIL NIL 1 2000
4	41	0	281000	315	NIL	1	1000	1	281000; 315 NIL 1 1 1000
4	42	1	282040	0	NIL	1	1000	1	282040; 0 NIL 1 1 1000

Table 3. Critères filmiques utilisés pour analyser les plans-séquences de *Bout-à-bout*. La partie en jaune à droite montre les points d'analyse, avec une date (en ms dans le fichier son) et les valeurs des cinq critères.

On peut donc créer un scénario qui générerait une séquence où, par exemple, on verrait uniquement des métros qui vont de gauche à droite. On pourra exiger de ce scénario qu'il conserve au maximum la continuité des plans d'origine (ou le contraire).

5.2.1. Projet pédagogique

Ces caractéristiques purement visuelles ont été expérimentées dans le cadre d'un enseignement de master à la faculté des arts de l'université de Strasbourg. Une maquette a été réalisée par Quentin Barrois, qui tire parti de l'extrême répétitivité formelle des films d'Ozu.

5.2.2. L'exemple d'Ozu

Les extensions vidéo de Omax-DYCI2 invitent tout naturellement à se pencher sur des cinématographies où la répétition a une place importante, telle celle de Yasujirô Ozu. Au cours de sa carrière, le cinéaste utilise de plus en plus de motifs similaires, jusqu'aux visages de ses acteurs qui migrent de film en film. Mais la structure répétitive organise les films en eux-mêmes et il est possible de dresser une liste des scènes types qui reviennent plusieurs fois au cours de la narration. Avec cette idée comme point de départ, il est aisé d'imaginer un projet idéal d'un film « ozuien », reconstruit grâce au logiciel à partir de différents films d'Ozu.

Tout d'abord, l'évidence pousse à se servir des acteurs habituels d'Ozu pour fondre ses différents films en un, comme nous y invitent les textes qui parlent de *trilogie Kihachi* ou de *trilogie Noriko* pour les films où un même acteur incarne un personnage portant le même nom (respectivement Takeshi Sakamoto et Setsuko Hara) [4, p. 41 et 80]. Ce sont loin d'être des cas isolés, et Chishu Ryû, Kuniko Miyake, Shin Saburi ou Haruko Sugimura sont également des noms qui reviennent à maintes reprises. Il s'agirait alors d'écrire en fondant tous leurs personnages, de donner à voir avec plus d'évidence cette confusion qui s'opère parfois entre les films, entre les rôles qui se ressemblent et, ici, s'assembleraient.

Pour permettre cela, il s'agit de garder en tête qu'« il n'y a nullement, chez Ozu, du remarquable *et* de l'ordinaire, des situations limites *et* des situations banales, les unes ayant un effet en venant s'insinuer aux autres » [13, p. 24]. Si l'on prend au sérieux cette idée, il n'est pas nécessaire de conserver le déroulement dramatique de tel ou tel film, l'union des films ne pouvant mener qu'à une suite de *banalités remarquables*, dont la présence est d'ailleurs particulièrement évidente dans les dialogues. Il serait exemplaire de pouvoir obtenir une séquence où les répliques se suivent sans créer d'enjeu, où les descriptions météorologiques sont des arguments aussi valables que l'origine sociale d'un prétendant au mariage.

Ainsi, un projet idéal sur Ozu tiendrait compte du fait que son cinéma est une image du temps qui s'écoule en cycle, n'ayant à ce titre ni vraiment de début ni de fin, et pouvant confondre toutes les incarnations d'un même type de rôle, tous les sujets abordés, en une séquence potentiellement infinie où apparaîtraient d'une part les lois qui régissent son univers, d'autre part « tout un réseau "intertextuel" d'échos à d'autres images et d'autres situations – dans un même film et d'un film à l'autre » [15, p. 61], qui est la source de la condensation des sentiments complexes des personnages.

Il faut terminer en évoquant le fait qu'il existe un nombre considérable d'autres options pour reconstituer Ozu, tant son œuvre est ponctuée de règles qui lui sont

propres. Ce serait par exemple le cas d'une réécriture qui ne prendrait en compte que le changement d'angle d'un plan à l'autre, selon les règles déterminées par David Bordwell [8, p. 90-94]. De toute manière, l'exercice rendu possible par le logiciel invite à rechercher d'autres règles de ce type en se noyant dans le « "trop plein", allant même parfois jusqu'à friser la saturation suffocante des signes » [15, p. 13], pour essayer de chercher des règles de construction dans les motifs des panneaux coulissants en papier de riz, dans les couleurs des chemises, ou dans le nombre et la forme des bouteilles...

6. CONCLUSION

L'architecture *OMax-DYCI2* permet de fascinantes expérimentations avec l'image.

Les créations visuelles interactives dans lesquelles le son est moteur ne sont pas si fréquentes. Par nature, l'utilisation de la synthèse concaténative favorise l'utilisation de la vidéo.

Des projets comme *Paris bout-à-bout* donnent une idée de tout le parti que l'on peut tirer de ces outils dans un contexte artistique.

Les diverses stratégies d'improvisation mises en œuvre dans *OMax* et *DYCI2* sont en effet intéressantes dans leurs applications visuelles. La navigation libre en fonction d'un contexte semblable, la navigation par scénario et l'écoute de l'improvisation pour guider le scénario, voire le créer, sont autant de manières d'explorer les relations entre image et son. Elles invitent autant à entendre l'image qu'à voir le son.

La possibilité d'enchaîner ou de boucler des processus permet d'envisager des dispositifs où la séquence des images est pilotée par une structure sonore, laquelle ne serait pas forcément entendue.

Enfin, la généralité de l'idée de scénario et la possibilité d'utiliser des descripteurs pour une construction automatique des mémoires laissent augurer une utilisation de cet environnement sur des critères purement visuels, comme l'a montré l'expérience sur les films d'Ozu. À terme, on pourrait faire appel à des techniques de reconnaissance d'image pour en déduire les analyses et les scénarios.

7. RÉFÉRENCES

[1] Allauzen, C., Crochemore, M., Raffinot, M. "Factor Oracle : A New Structure for Pattern Matching », *Proceedings of SOFSEM'99 : Theory and Practice of Informatics*, Pavelka, J., Tel, G., Bartosek, M. (dir.). Springer, Berlin, 1999, p. 291-306. [Lecture Notes in Computer Science, vol. 1725.]

[2] Assayag, G., Bloch, G., Chemillier, M. « OMax-Ofon », Sound and Music Computing (SMC) Conference, Marseille, 2006.

[3] Assayag, G., Bloch, G. « Navigating the Oracle : A Heuristic Approach », Proceedings of the International Computer Music Conference, Copenhague, Danemark, 2007.

[4] Beth, S. *L'impuissance du cinéma : une étude des films d'Ozu*. Presses universitaires de Strasbourg, Strasbourg, 2018.

[5] Bloch, G., Chabot, X., Dannenberg, R. « A Workstation in Live Performance: Composed Improvisation », Proceedings of the International Computer Music Conference, La Haye, Pays-Bas, 1986.

[6] Bloch, G., Dubnov, S., Assayag, G. « Introducing Video Features and Spectral Descriptors in The OMax Improvisation System », Proceedings of the International Computer Music Conference, Belfast, Irlande, 2008.

[7] Bloch, G., Nika, J. « Edit Piaf, Billie Holiday, and Elisabeth Schwarzkopf Making Music Together », *Exploring Transdisciplinarity in Art and Sciences*, Kapoula, Z., Volle, E., Renoult, J., Andreatta, M. (dir.). Springer, Berlin, 2018, p. 275-299.

[8] Bordwell, D., *Ozu and the Poetics of Cinema*. Princeton University Press, Princeton (NJ), 1988.

[9] Carsault, T., McLeod, A., Esling, P., Nika, J., Nakamura, E., Yoshii, K. « Multi-Step Chord Sequence Prediction Based on Aggregated Multi-Scale Encoder-Decoder Network », IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP), Pittsburgh, USA, 2019.

[10] Carsault, T., Esling, P., Nika, J. « Using Musical Relationships Between Chord Labels in Automatic Chord Extraction Tasks », International Society for Music Information Retrieval Conference (ISMIR), Paris, 2018.

[11] Chadabe, J. « Some Reflections on the Nature of the Landscape within which Computer Music Systems are Designed », *Computer Music Journal* 1/3 (1977), p. 5-11.

[12] Deguernel, K., Vincent, E., Assayag, G. « Using Multidimensional Sequences for Improvisation in the OMax Paradigm », Proceedings of the Sound and Music Computing Conference, Hambourg, Allemagne, 2016, p. 117-122.

[13] Deleuze, G. *Cinéma 1 : l'image-temps*. Paris, Minuit, 1983.

[14] Deleuze, G. *Différence et répétition*. PUF, Paris, 2011. [1re éd. 1968.]

[15] Doganis, B. *Le silence dans le cinéma d'Ozu : polyphonies des sens et du sens*. L'Harmattan, Paris, 2005.

[16] Dubnov, S., Assayag, G. « Universal Prediction Applied to Stylistic Music Generation », *Mathematics and Music : A Diderot Mathematical Forum*, Assayag, G., Feichtinger, H. G., Rodrigues J. F. (dir.). Springer, Berlin, 2002, p. 147-158.

[17] François, A. R., Schankler, I., Chew, E. « Mimi4x : An Interactive Audio-visual Installation for High-level Structural Improvisation », *International Journal of Arts and Technology* 6/2 (2013), p. 138-151.

[18] Grabócz, M., *Musique, narrativité, signification*. L'Harmattan, Paris, 2009. [Préface de Charles Rosen.]

[19] Nika, J. « Guiding Human-Computer Music Improvisation : Introducing Authoring and Control with Temporal Scenarios », thèse de doctorat, sous la dir. de G. Assayag et M. Chemillier, Université Pierre-et-Marie-Curie (Paris 6), 2016.

[20] Nika, J., Déguernel, A., Chemla-Romeu-Santos, A., Vincent, E., Assayag, G. « DYCI2 Agents : Merging the "Free", "Reactive" and "Scenario-based" Music Generation Paradigms », Proceedings of the 43rd International Computer Music Conference, Shanghai, Chine, 2017, p. 227-232.

[21] Sellin, H. *Always too Soon*, CD Cristal Records, 2019.

[22] Surges, G., Dubnov, S., « Feature Selection and Composition Using PyOracle », Proceedings of the Artificial Intelligence and Interactive Digital Entertainment Conference, Boston, USA, 2013, p. 114-121.

Texte édité par Corentin Guichaoua