EARTOY : INTERACTIONS LUDIQUES PAR L'AUDITION

Isabelle Viaud-Delmon
CNRS UMR 7593
ivd@ext.jussieu.fr

Jean Bresson
IRCAM
jean.bresson@ircam.fr

François Pachet
SONY CSL
pachet@csl.sony.fr

Frédéric Bevilacqua
IRCAM
frederic.bevilacqua@ircam.fr

Pierre Roy
SONY CSL
roy@csl.sony.fr

Olivier Warusfel
IRCAM
olivier.warusfel@ircam.fr

RÉSUMÉ

Bien que le monde du son ait évolué de concert avec les progrès technologiques, le monde des images semble dominer notre perception. Par exemple, la réalité virtuelle (RV) est bien souvent la métaphore inconsciente d'une réalité virtuelle visuelle, et les applications de RV, qu'elles relèvent du jeu, de l'apprentissage ou même de l'art, s'adressent principalement à la vision. On sait l'importance des notions d'espace et d'interaction pour promouvoir la sensation de présence dans des environnements virtuels. Cependant, peu de travaux s'attachent à exploiter le potentiel fourni par les stimulations virtuelles s'adressant au système auditif de l'utilisateur. Nous présentons ici le projet EarToy, dont l'objectif vise à mettre à jour les liens existant entre audition et action, afin de construire des systèmes immersifs et interactifs basés sur la perception auditive.

1. INTRODUCTION

Les images dominent notre perception du monde. Pourtant, l'audition est la seule modalité sensorielle mettant l'Homme en relation avec l'ensemble de l'espace qui l'entoure. Elle se distingue de toutes les autres modalités sensorielles (vision, toucher) dont la disposition des organes récepteurs sur le corps ne permet la captation des stimuli que dans un champ limité.

Par ailleurs, si les dispositifs de reproduction restituant un champ sonore cohérent spatialement restent encore confinés au monde du laboratoire, ils sont paradoxalement en avance sur leurs équivalents visuels : harnachement moins lourd, diffusion sans fil, liberté de mouvement sur de grandes distances. Libéré de la contrainte du regard porté vers un écran, l'utilisateur d'un système de reproduction sonore peut explorer sans restriction les relations qui le lient à l'espace immersif.

Nous nous intéressons dans le projet EarToy à l'audition dans ses dimensions spatiales et à son intégration avec les modalités sensorielles non-visuelles afin d'augmenter la potentialité immersive des environnements virtuels. L'objectif principal est d'étudier l'interaction audition/corps/espace dans les dispositifs de reproduction sonore dits immersifs, de promouvoir les applications interactives accordant la primauté à la modalité sensorielle auditive et de développer de nouveaux paradigmes d'interaction avec un contenu sonore. La conception de tels environnements interactifs centrés sur la modalité auditive soulève des questions scientifiques stimulantes et présente un potentiel d'applications innovantes dans de nombreux domaines. Nous nous intéressons en particulier aux domaines du jeu, de la recherche expérimentale sur le comportement humain, et des installations artistiques.

Ce projet consiste à exploiter les technologies de restitution audio 3D et de capture du mouvement dans leurs développements les plus récents, et à explorer le potentiel de la modalité d'interaction audition/corps/espace à travers des scénarios d'usage ludiques s'appliquant à la recherche expérimentale et au jeu musical. Nous souhaitons démontrer avec ce projet les potentialités offertes par les stimulations auditives pour la RV en explorant trois axes : 1) la création de mondes virtuels auditifs évolutifs, 2) la création d'un avatar sonore, et 3) la mise en place de nouvelles modalités d'interaction entre l'utilisateur et le monde virtuel.

2. REALITE VIRTUELLE AUDITIVE

Dans les mondes virtuels, les tâches de navigation sont généralement confiées à la vision associée à des dispositifs gestuels auxquels sont délégués les contrôles de déplacement et de sélection. Dans le cadre de la RV auditive telle que nous l'envisageons, l'utilisateur est placé dans une situation de navigation par l'audition et par l'action. Il n'a à manipuler aucun périphérique de contrôle (souris, joystick ou autre). Ce sont ses mouvements propres qui sont analysés en temps réel de sorte à autoriser une interaction naturelle. Les avantages attendus d'une telle interaction avec un monde auditif sont de reposer sur des modalités sensorielles éminemment tridimensionnelles. Par ailleurs, notre connaissance et notre représentation spatiale sont intimement liés aux processus d'apprentissage

engageant le corps. C'est la maîtrise conjointe de ces deux modalités, audition et action, qui nous permettra de synthétiser et contrôler des espaces sonores interactifs.

Un de nos objectifs est d'élaborer différents modèles d'interaction régissant la mise à jour du monde sonore en fonction des mouvements de l'utilisateur. Cette phase peut s'appuyer sur la constitution d'un répertoire d'interactions reposant sur une description symbolique des règles d'association entre une classe de gestes et des procédés d'articulation, de transformation et de spatialisation d'un corpus sonore. Une autre approche consiste à inférer de manière implicite ces règles d'association à partir de l'analyse du comportement de l'utilisateur selon un principe d'interaction réflexive [1]. La pertinence au niveau perceptif et cognitif de ces modèles d'interaction sera évaluée expérimentalement.

Nous proposons également d'élaborer le concept d'avatar sonore, manifestation auditive de la présence de l'utilisateur dans le monde virtuel. La notion d'avatar sonore nécessite la prise en compte de la nature particulière du son pour concevoir les caractéristiques de sa manifestation sonore en fonction du type de tâche effectuée (sélection, focus, loupe, déplacement, transformation, etc.) et du contexte dans lequel elle s'insère. Par ailleurs, la conception de l'avatar doit résulter d'un compromis entre, d'une part, la facilité de son repérage et de son interprétation et, d'autre part, la minimisation de sa perturbation et de son masquage du monde sonore (surcharge auditive).

3. APPLICATIONS

Les concepts et techniques développés sont évalués dans le cadre d'applications pilotes. Une des premières applications est l'étude comportementale de l'Homme dans des situations d'immersion et d'interaction avec un monde auditif. Un autre domaine concerne l'accès interactif à la musique.

3.1. Etude de la perception auditive spatiale et de ses liens avec le corps

Les attributs spatiaux de la perception auditive et les signaux qui accompagnent ces attributs constituent ce qui est appelé l'audition spatiale [2]. La perception auditive est une des voies importantes, avec la vision, par lesquelles l'Homme accède à la connaissance de l'espace. Elle apporte, contrairement à la vision, des informations sur l'entité de l'espace environnant l'individu sans se restreindre à l'espace frontal. Malgré cet état de fait, la modalité auditive est encore peu étudiée dans ses dimensions spatiales. Le développement de tels dispositifs invite à la mise en place d'une démarche expérimentale sur l'audition spatiale dans un contexte écologique. Une telle démarche permet de plus de valider perceptivement les développements technologiques.

Les nouvelles formes de jeu, par exemple, offrent une plus grande richesse d'interactions sensorielles et nécessitent de la part de l'utilisateur l'acquisition de compétences d'intégration sensorielle inhabituelles par rapport à son cadre de vie quotidien. L'augmentation de l'interactivité dans les jeux, ainsi que dans les installations artistiques, va nécessairement requérir que l'audition soit un élément clé du développement de nouvelles formes d'interaction sensorielle chez l'utilisateur. Ce nouveau type d'interaction vient ainsi questionner les frontières existantes entre jeu, applications artistiques, et expérimentations scientifiques. Il représente un enjeu à la fois pour la technologie logicielle et l'étude de la perception humaine.

Nous avons commencé une série d'expériences calquées sur un paradigme classique en neurosciences, le «Morris Watermaze» [3]. Le dispositif utilisé permet aux sujets de l'expérience, équipés de marqueurs de position de la tête, de déambuler librement dans une zone de 30 m2. Les sujets sont équipés d'un casque audio sans fil délivrant un paysage sonore mis à jour en temps réel en fonction de leurs mouvements dans l'espace.

La tâche des sujets est de retrouver une cible sonore ne se déclenchant que lorsqu'ils ont atteint une position précise dans l'espace. Pour réaliser cette tâche, ils ne disposent pas d'information visuelle, car ils sont dans le noir complet. L'espace virtuel dans lequel ils sont immergés est purement sonore. Ainsi, la réussite à la tâche repose uniquement sur les capacités du sujet à traiter les informations idiothétiques[1] générées au cours de la déambulation en relation avec les informations auditives fournies par l'espace sonore. Ce type de situation fait instinctivement penser à l'étude de la perception chez les sujets aveugles. Cependant, cette condition de déprivation visuelle permet de mettre en évidence que chez l'Homme normo-voyant, les indices auditifs participent automatiquement au système de repérage spatial. Chez l'Homme, la prédominance du canal visuel est telle que pour mettre en évidence le rôle d'une autre modalité sensorielle dans le système de repérage spatial, il est nécessaire d'en isoler la contribution.

Les résultats de ces expériences indiquent que l'Homme normo-voyant est capable de construire une carte cognitive spatiale sur la seule base d'indices idiothétiques et auditifs, et que cette construction intervient en l'absence de construction d'un équivalent visuel. Nous pouvons ainsi d'ores et déjà suggérer que les situations dans lesquelles l'utilisateur navigue par l'audition et par l'action obtiennent l'agrément perceptif.

[1] Ensemble des informations engendrées par les mouvements du corps, regroupant les informations vestibulaires (organe de l'équilibre situé dans l'oreille interne, captant les accélérations linéaires et angulaires de la tête) et proprioceptives (informations liées aux muscles, tendons, articulations et capteurs de pression répartis sur le corps).

3.2. Musique interactive

L'émergence des technologies de codage par le contenu et l'association de meta-données aux flux musicaux enregistrés ou transmis permet d'envisager de nouveaux modes d'accès et de diffusion musicale, ainsi que de nouveaux usages du « faire de la musique ». La navigation dans un catalogue musical, ou l'écoute interactive, est rendue possible par l'exploitation de descripteurs sonores éventuellement extraits à la volée. En particulier, il est possible de développer des mécanismes d'interaction avec le contenu dépassant le cadre de contrôles élémentaires agissant sur les caractéristiques du signal audio (niveau sonore, balance, égalisation).

Nous envisageons d'élaborer un jeu musical interactif, permettant au joueur de produire de la musique tout en navigant dans une base de morceaux musicaux dont il sélectionne et recombine certains extraits préférés.

Les règles d'association entre les gestes et le retour sonore pourront être établies à différents niveaux de description du contenu sonore, depuis le niveau signal jusqu'à des données d'ordre symbolique. Si l'interaction agit sur un corpus musical raisonné (catalogue indexé), les tâches d'exploration et d'écoute s'imbriquent dans une seule et même activité de type musicing [4]. L'évolution du monde induite par les actions de l'utilisateur pourra par conséquent procéder de différents niveaux d'articulation du discours musical en intervenant sur la structure temporelle, le démixage, la substitution de timbres ou d'instruments, etc.

4. TECHNOLOGIES DE REPRODUCTION SONORE SPATIALE, IMMERSIVE ET INTERACTIVE

Des dispositifs de reproduction sont aujourd'hui disponibles pour restituer un champ sonore cohérent spatialement, c'est-à-dire où les indices acoustiques associés à la perception auditive spatiale varient de manière congruente avec les déplacements de l'auditeur. Les deux seules techniques répondant à cette contrainte sont les techniques holophoniques pour les applications collectives et la technique binaurale pour les applications individuelles. Cette dernière technique implique un suivi dynamique de l'utilisateur du système afin de préserver la cohérence du champ sonore en fonction de ses déplacements ou mouvements.

4.1. Technique binaurale

Longtemps confinée au domaine du laboratoire, la technique binaurale a aujourd'hui atteint un degré de maturité qui permet son utilisation dans des contextes plus grand public comme le jeu. Cette technique de diffusion 3D pour casque et basée sur la synthèse des fonctions de transfert d'oreille (HRTF – *Head Related Transfer Function*) et permet une reconstruction fine et exhaustive des indices acoustiques responsables de la localisation auditive spatiale. Des solutions opérationnelles intégrées dans un environnement de spatialisation temps réel ont déjà été développées [5][6]. À l'instar des casques de vision stéréoscopique, le couplage de la technique binaurale avec un dispositif de suivi de la position et de l'orientation de la tête permet d'asservir la scène sonore diffusée aux mouvements de l'auditeur.

La mise en jeu de cette boucle de rétroaction entre proprioception et audition est déterminante pour l'adhésion de l'auditeur à la scène sonore virtuelle qui lui est présentée puisqu'elle permet de garantir la congruence entre les variations des indices acoustiques binauraux et ses mouvements propres. Les lois de rétroaction proprioceptive et auditive étant définies entièrement au niveau logiciel, celles-ci peuvent être choisies de sorte à créer des mondes sonores régis par des lois non réalistes. On peut alors solliciter la perception de façon insolite et motivante pour les applications artistiques. Cette technique et ces principes sont utilisés pour mener les expériences de validation perceptive et cognitive, et pour le démonstrateur de jeu musical.

4.2. Système de repérage spatial, capture et analyse du geste

Différents dispositifs de caméras permettant de suivre la position d'objet ou de personnes dans l'espace sont aujourd'hui disponibles et couramment utilisés pour la réalité virtuelle. Associés à ces matériels de captation, plusieurs logiciels sont apparus ces dernières années et proposent divers types d'analyse vidéo. Par exemple, EyesWeb [7], développé par l'université de Gênes, est un logiciel qui permet grâce à une programmation graphique d'effectuer en temps réel une chaîne de traitements d'image et d'analyse gestuelle. Il contient des modules standard d'extraction de la silhouette, suivi de blobs ou de couleurs. Des modules spécifiques, écrits en C++, peuvent être ajoutés à la librairie standard.

Dans EarToy, nous utilisons deux types de systèmes de captation différents, suivant les phases du projet. Un système de caméras infra-rouges couplé à des marqueurs passifs (de type AR-Tracking) est utilisé pour les expériences de validation scientifique, nécessitant des hautes performances en termes de précision spatiale, latence et zone de couverture. Pour le scénario applicatif, en revanche, l'utilisation de deux webcams est visé de sorte à préfigurer un matériel adapté aux applications grand public.

Les données gestuelles sont ensuite traitées en temps réel grâce à des logiciels tels que Max/MSP et la librairie d'analyse de geste MnM [8].

5. CONTEXTE

Le projet EarToy est issu de travaux précurseurs menés entre autres dans les domaines de la réalité virtuelle augmentée et de la sonification.

5.1. Le Projet LISTEN

L'objet de LISTEN était d'explorer la notion de réalité sonore augmentée en l'appliquant au contexte muséographique [9]. Dans LISTEN, l'utilisateur, équipé d'un casque audio sans fil et repéré dans l'espace par un dispositif de caméras, parcourt un lieu réel auquel est superposé un univers sonore constitué de divers éléments, motifs musicaux ou messages didactiques, répartis dans l'espace et éventuellement associés à la présence physique d'un objet (un tableau, une sculpture). L'interaction est fondée sur une tâche de navigation de l'auditeur pour explorer l'organisation spatiale des éléments sonores. L'interaction peut également être enrichie en insérant des règles de partitionnement de l'espace (zones de déclenchement d'événements) et de dépendance temporelle (analyse du parcours de l'auditeur), incitant dès lors l'auditeur à jouer avec le contenu et tenter de l'influencer par son comportement. Ainsi, le contenu délivré ne se borne pas à une procédure déclarative et verbalisée, mais s'accompagne d'informations ou d'impressions s'adressant à différents niveaux de perception. En retour, l'espace et le temps perçus subjectivement par le visiteur sont étroitement liés aux éléments sonores qui les ont habités et aux processus de segmentation temporelle qui les ont animés.

En partant de l'expérience du projet LISTEN, plusieurs pistes d'approfondissement sont dégagées et servent d'objectifs initiaux pour EarToy. L'accès aux paramètres posturaux permettra d'augmenter considérablement la finesse d'analyse du comportement de l'auditeur et par conséquent le vocabulaire d'interaction (rythme de la marche, posture d'attente,…).

5.2. Sonification - Audio-Only gaming

Le concept d'avatar sonore proposé par le projet EarToy renvoie aux travaux sur la sonification qui visent à exploiter la modalité auditive en traduisant sous forme sonore une information visuelle afin de limiter la surcharge d'informations fournie par les interfaces graphiques. Dans le cas de flux de données abstraites, l'observation par l'audition est plus performante que la vision pour le repérage de structures temporelles.

Plusieurs recherches récentes, notamment en Allemagne [10][11], font également mention de jeux s'adressant exclusivement à la modalité auditive (audio-gaming only). Cette démarche est encore embryonnaire, et les prototypes consistent généralement à transposer dans le domaine sonore des principes de jeux s'adressant traditionnellement à la vision (jeux d'adresse, jeux d'aventures). Les récents articles scientifiques parus témoignent du caractère captivant de ce type de jeu pour les premiers utilisateurs et confirment l'intérêt de consacrer un travail de recherche et de développement autour de ce concept.

6. ORGANISATION GLOBALE DU PROJET

L'exploration de la modalité d'interaction auditive / idiothétique repose sur la mise en place d'une architecture matérielle et logicielle, permettant à un utilisateur de se déplacer dans un espace réel augmenté par la superposition d'une scène sonore, diffusée par un casque sans fil. L'interaction auditive/idiothétique réside dans l'asservissement du contenu et de l'organisation spatiale de la scène sonore aux mouvements de l'utilisateur/auditeur. Les mouvements sont saisis par des dispositifs de repérage spatial et de capture posturale du sujet.

Sur le plan technologique, le projet EarToy s'appuie avant tout sur des systèmes de reproduction sonores et de capture maîtrisés de sorte à minimiser les temps de développement bas niveau au profit de l'exploration scientifique de la modalité d'interaction auditive et idiothétique et du développement d'outils d'écriture de l'interaction.

6.1. Interactions élémentaires – Cognition Spatiale

Nous entendons procéder par invention de différentes « mises en situation » élémentaires ayant pour but de révéler simultanément les propriétés de l'interaction auditive et idiothétique sur le plan scientifique (cognition spatiale) et les principaux composants logiciels nécessaires à leur mise en œuvre. Le travail expérimental en neurosciences constituera l'un des axes majeurs du projet proposé puisqu'il doit nourrir le travail de recherche et de développement technologique et évaluer les facultés d'appropriation de la modalité audition/corps/espace virtuel par les sujets.

6.2. Algorithmes de traitement de la captation spatiale et posturale

L'exploration de différentes modalités d'interaction nécessite la création d'algorithmes spécifiques de traitement de la posture et le développement de modules de traitement et d'encodage des signaux issus de la capture spatiale. Ces données sont communiquées au synthétiseur sonore pour établir un retour auditif en temps-réel et avec une latence minimale. Les interactions de plus haut-niveau pourront nécessiter une extraction préalable de formes de sorte à fournir aux outils d'écriture des données dans un format plus symbolique.

6.3. Module technique pour l'interaction et la synthèse sonore

La gestion de l'application globale comprend les tâches de construction de l'environnement sonore, de définition des interactions et de contrôle de leurs

évolutions au cours du déroulement de l'application. Les modules d'interaction peuvent intervenir à deux niveaux parallèles selon le grain temporel nécessaire. Par exemple, les paramètres de spatialisation des sources sonores composant la scène virtuelle doivent pouvoir être rafraîchis en temps réel avec une latence inférieure au seuil de perception ce qui invite à traiter l'interaction le plus directement possible entre les modules de capture spatiale et les modules de synthèse sonore. D'autres types d'interaction peuvent en revanche être traitées sans contrainte stricte de réactivité et nécessiter au contraire de travailler sur un corpus de données symboliques comme par exemple dans le cas d'un processus d'apprentissage du comportement de l'utilisateur. Ces deux niveaux peuvent également communiquer de manière à modifier les lois d'interaction temps-réel au cours de l'application.

7. CONCLUSION

L'idée centrale sous-jacente au projet EarToy est d'exploiter un principe d'interaction audition/corps/espace rendu accessible par le recours aux techniques dérivées de la réalité virtuelle. Notre démarche est radicale en ce qu'elle consiste à modifier les modalités d'interactions usuelles en plaçant l'utilisateur dans une situation d'exploration accordant la primauté à l'audition et l'action, voire excluant la vision. Cette démarche fait l'objet de recherches et de développements dans les communautés des neurosciences et de la réalité virtuelle, notamment du jeu, mais place le domaine musical dans une position naturellement privilégiée pour aborder ces questions.

8. REMERCIEMENTS

Le projet EarToy est financé par l'ANR dans le cadre du RIAM.

9. REFERENCES

[1] Pachet, F., Addessi, A.R. "When Children Reflect on Their Playing Style: The Continuator." *ACM Computers in Entertainment*, 2004, 1:14.

[2] Blauert, J. *Spatial Hearing*. Revised ed. MIT Press, Cambridge, MA, 1997.

[3] Morris, RGM. "Spatial localization does not require the presence of local cues", *Learning and motivation* 1981; 12:239–60.

[4] Zils, A., Pachet, F. "Musical Mosaicing", *Proceedings of DAFX 01*, December 2001, University of Limerick.

[5] Jot, J.-M. "Real-time spatial processing of sounds for music, multimedia and interactive human-computer interfaces", *Multimedia Systems* 1999; 7:55–69.

[6] Blum, A., Katz, B., Warusfel, O. "Eliciting adaptation to non-individual HRTF spectral cues with multi-modal training", *Proceedings of CFA-DAGA 04*, Strasbourg, 2004.

[7] Camurri, A., Hashimoto, S., Ricchetti, M., Trocca, R., Suzuki, K., Volpe, G. "EyesWeb – Toward Gesture and Affect Recognition in Interactive Dance and Music Systems", *Computer Music Journal*, 2000; 24:57–69.

[8] Bevilacqua, F., Muller, R., Schnell, N. "MnM: a Max/MSP mapping toolbox", *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME 05),* p. 85-88, 2005.

[9] Eckel, G. "Immersive Audio-Augmented Environments – The LISTEN Project", *Proceedings of the 5th Int. Conf. on Information Visualization (IV2001)*, IEEE Computer Society Press, 2001.

[10] Beilharz, K. "Wireless gesture controllers to affect information sonification", *Proceedings of ICAD 05*, 2005.

[11] Röber, N., Masuch, M. "Leaving the screen – new perspectives in audio-only gaming", *Proceedings of ICAD 05*, 2005.