Le son 3D dans toutes ses dimensions

R. Nicol, J. Daniel, M. Emerit, G. Pallone, D. Virette, N. Chetry, P. Guillon, S. Bertet
France Telecom R&D Orange Labs
4, rue du Clos Courtel
BP 91226
35512 Cesson Sévigné
Tel. : 02 99 12 41 11
E-mail : rozenn.nicol@orange-ftgroup.com

La vidéo 3D est annoncée comme la prochaine évolution du monde de l'image après la Haute Définition (HD). Qu'en est-il pour le son? En fait, sans que nous en soyons véritablement conscients, nous sommes déjà plongés dans l'ère du son 3D. Aujourd'hui le son multicanal 5.1 a envahi les salles de cinéma et nos salons. Le son 3D est aussi présent dans les jeux 3D interactifs sur PC ou consoles. Mais ce n'est là qu'un petit échantillon de ce que le son 3D peut nous offrir aujourd'hui et de ce qu'il nous réserve pour l'avenir…

Dans notre expérience naturelle d'écoute, nous percevons le monde sonore en 3D, c'est-à-dire que nous sommes capables d'identifier la position des sons à la fois en distance, en azimut et en élévation. Ce qui nous permet de localiser les sons, c'est d'abord le fait que nous avons deux oreilles ; les différences entre les sons captés au niveau des tympans gauche et droit renseignent le système auditif sur l'azimut de la scène sonore. Ces différences interaurales portent sur le temps d'arrivée (ITD pour Interaural Time Difference) et sur l'intensité des sons (ILD pour Interaural Level Difference). ITD et ILD sont les indices majeurs de la localisation dans le plan horizontal. Le second organe déterminant pour la localisation des sons est le pavillon de l'oreille ; par un jeu de résonances et de diffractions, le pavillon contribue à modifier le timbre des sons en fonction de leur direction d'incidence. Ces indices spectraux sont exploités par le système auditif pour identifier l'élévation de la source sonore. D'autres éléments morphologiques (réflexions sur le torse et les épaules) viennent aussi compléter le rôle du pavillon.

Les technologies du son 3D visent à créer l'illusion de cette perception en 3D. Mais, en ajoutant une nouvelle dimension à la restitution du son, le son 3D permet surtout d'introduire une vraie rupture dans les applications audio. Par exemple, dans le domaine des télécommunications, il permet d'enrichir les services de communication de groupe. La flexibilité intrinsèque de la VoIP permet d'envisager la généralisation de communications interpersonnelles enrichies par la dimension spatiale. Des services professionnels comme la conférence audio ou la visioconférence pourront ainsi profiter de l'intelligibilité, du naturel et du confort apportés par un rendu sonore spatialisé des locuteurs. L'étape suivante est la «téléportation sonore» c'est-à-dire capter la scène sonore de façon si fidèle qu'une fois restituée, les auditeurs ont l'illusion d'être transportés sur les lieux de la prise de son. C'est la possibilité de faire partager à des amis distants des événements tels qu'un concert ou une fête de famille. Les techniques de captation audio 3D qui permettent cette immersion sonore 3D existent déjà ; il s'agit par exemple de la prise de son binaurale avec une tête artificielle. Les futures interfaces misent aussi sur le son 3D : avec la dimension spatiale sonore on peut construire un écran sonore qui vient compléter l'écran visuel en exploitant une structure spatiale similaire. C'est aussi une piste pour développer des interfaces dédiées aux aveugles.

Les technologies audio 3D sont-elles aujourd'hui à la hauteur de ces futurs enjeux ? Certes, le monde du son 3D est riche de technologies : stéréophonie, multicanal, binaural, holophonie (Wave Field Synthesis ou WFS), Ambisonics (Higher Order Ambisonics ou HOA), etc. Cette variété est d'abord un atout ; aucune technologie n'est idéale en soi, chaque technologie présente des avantages et des inconvénients dont on peut jouer pour définir la solution la mieux adaptée à un contexte donné d'application. Mais la pluralité des technologies audio 3D est aussi un frein, dans la mesure où elle contribue à multiplier les formats de contenus. Les créateurs (ingénieurs du son), les distributeurs (radio, télévision, cinéma, internet) et le grand public sont facilement noyés dans cette diversité. La principale difficulté réside dans le fait qu'un format de contenu est dédié en général à un système d'écoute spécifique : système stéréophonique, système 5.1, casque… Si l'on veut pouvoir accéder à tous les contenus proposés, il faudrait multiplier les systèmes d'écoute. Un enjeu fort pour le domaine de l'audio 3D est donc la conversion de formats afin de pouvoir adapter les contenus à la configuration d'écoute disponible. Ici le terme contenu est vu au sens large. Il englobe aussi bien la musique, les films, les jeux,

les messages multimédia, les contenus autoproduits, les signaux des interfaces hommes-machines que les contenus évanescents ou produits en temps réel dans le contexte de l'audioconférence, la visioconférence, la radio ou la télévision en direct.

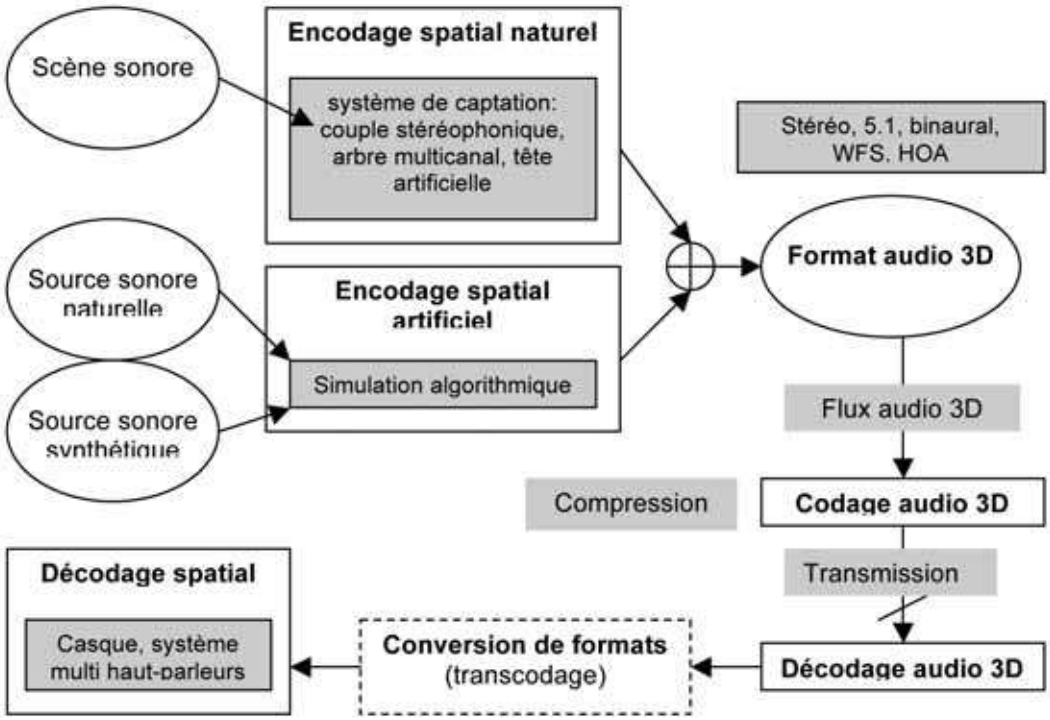


Fig. 1: Autour des formats audio 3D : encodage (captation naturelle ou artificielle) et décodage spatial, codage et décodage audio 3D, transcodage

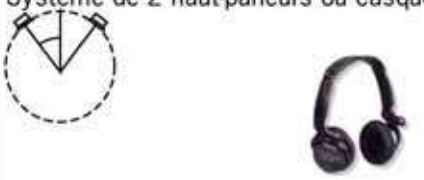
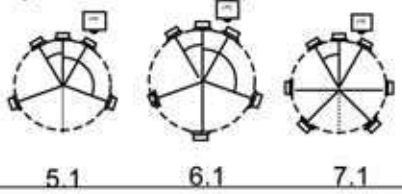
De ces questions, une notion fondamentale émerge : la notion de **format audio 3D**, avec les notions associées d'**encodage** et de **décodage** spatial, de **conversion** de formats et de **compression** audio 3D (Fig. 1). L'ensemble de ces notions sont aujourd'hui en cours de définition. Une question clef pour la création et le transport de contenus audio 3D est l'émergence d'un format universel permettant de s'affranchir des opérations de transcodage tout en garantissant une qualité optimale de spatialisation. Les concepts qui viennent d'être introduits seront illustrés dans la suite de l'article autour des différentes technologies audio 3D. Dans un premier temps, on s'intéressera aux technologies disponibles aujourd'hui ou que l'on trouvera d'ici quatre ans dans les applications grand public (stéréophonie et multicanal). On verra qu'il existe déjà pour ces formats des techniques de transcodage (upmix, downmix) et de codage audio 3D. La seconde partie présentera des technologies qui n'ont pas encore conquis le grand public ou qui sont en cours de développement dans les laboratoires de recherche (binaural, WFS, HOA), en précisant les problèmes qu'elles tentent de résoudre.

Le son multicanal

Pour le grand public (home-cinema, jeux sur PC…), les contenus audio 3D se déclinent principalement autour des formats multicanal (5.1, 6.1, 7.1, 10.2, 22.2). La stéréophonie n'en est qu'un cas particulier et peut être considérée comme le format 2.0 (voire 2.1) du multicanal. Ces formats sont des références établies au sein de la communauté audio professionnelle et y sont largement utilisés. Ils font aussi l'objet de normes standardisées, notamment au niveau de MPEG (Moving Picture Expert Group). L'ensemble des traitements intervenant tout au long de la chaîne audio, de l'encodage (captation naturelle ou encodage artificiel) à la restitution, en passant par la transmission (codage multicanal) et le transcodage (upmix, downmix), est relativement bien maîtrisé pour ces formats.

Le(s) format(s) multicanal

La stéréophonie est le premier format de spatialisation sonore. Il utilise deux canaux qui alimentent chacun un haut-parleur. La position des sources sonores est contrôlée par des différences de niveau ou ICLD (Inter-Channel Level Differences), et/ou des différences de temps ou ICTD (Inter-Channel Time Differences) entre ces canaux (Tableau 1). Techniquement le format multicanal n'est qu'une généralisation des procédés stéréophoniques sur plus de 2 canaux. Les contenus multicanal se sont généralisés avec l'utilisation grand public des DVD-Vidéo et autres supports récents (DVD-Audio, HD-DVD, Blue-Ray…), mais aussi des jeux vidéo. Le principe du système multicanal consiste à répartir des haut-parleurs autour de l'auditeur afin de le plonger dans la scène sonore (sensation d'immersion), et dans le cas du cinéma pour couvrir une plus large zone d'écoute. Le format le plus répandu est le 5.1, mais d'autres formats, comme le 3.0, le 4.0, le 6.1, le 7.1, le 22.2

	Stéréophonie	Multicanal
Système de captation naturelle	*Encodage basé ICLD* (couples stéréophoniques à capsules coïncidentes): stereosonic, M-S, X-Y Avantages: compatibilité monophonique, précision de localisation	Encodage hybride d'ICTD et d'ITLD (arbres multicanal): multiple AB, Multichannel Microphone Array Design, INA-5, Fukada-Tree, OCT Surround, KFM-Surround [1]
	Encodage basé ICTD (couples stéréophoniques à capsules non-coïncidentes): AB ou AB-ORTF Avantage: qualité de l'impression spatiale	
Encodage spatial artificiel	Panoramique d'intensité ou intensity panning: loi sinus [2] et la loi tangente [3].	Lois de panning multicanal
Décodage spatial	Système de 2 haut-parleurs ou casque 	Systèmes multi haut-parleurs 

Tabl. 1: Encodage et décodage pour les formats multicanal

existent également. Au cinéma, l'usage est d'utiliser le canal central pour la voix, les canaux gauche/droit pour la musique, et les canaux arrière pour les effets sonores. Cela permet une certaine stabilité des scènes frontales. Pour les contenus musicaux, les haut-parleurs frontaux sont utilisés pour la musique, et les haut-parleurs arrière pour diffuser l'ambiance de la salle (réflexions arrière, présence du public…). Cependant, pour la musique, le format stéréophonique est encore le plus souvent préféré. La principale limitation des formats multicanal est la taille de la zone d'écoute qui se restreint à un point ou *sweet-point*. Même s'il est toujours possible d'augmenter le nombre de canaux, les formats multicanal restent limités dans leur principe par une spatialisation dédiée au plan horizontal et focalisée sur la zone frontale. La plupart des formats actuels ne permettent pas de contrôler la position des sources virtuelles en élévation. De plus, l'effet de profondeur est mal contrôlé.

Conversion de format

Lorsqu'on écoute un contenu multicanal 5.1 sur un système stéréophonique, on perd l'immersion et les effets introduits par les canaux arrière, ainsi que la stabilité de l'image sonore frontale apportée par le haut-parleur central. De même, quand on écoute un contenu multicanal 5.1 au casque, la scène sonore reste enfermée entre les deux oreilles. Enfin, quand on écoute un contenu stéréophonique sur une paire d'enceintes très proches, on perd l'effet stéréophonique. Les techniques de conversion de formats visent à réduire ou éliminer ces inconvénients. Face à la multiplicité des formats, un premier objectif est l'**adaptation** des contenus au système de restitution, étant donné que les formats de représentation, de transmission et de codage ne correspondent pas nécessairement à la configuration du système de reproduction et qu'il n'est en général pas envisageable pour un utilisateur d'avoir à disposition plusieurs systèmes de restitution. Une seconde finalité est l'**enrichissement** des contenus, afin de rajouter une dimension et de favoriser l'immersion dans la scène sonore. On rencontre aussi des traitements d'élargissement stéréophonique sur enceintes (pour donner l'impression que le son sort de deux haut-parleurs virtuels qui seraient plus écartés que les haut-parleurs réels) ou sur casque (pour externaliser les sources sonores en simulant des haut-parleurs virtuels par synthèse binaurale). Les outils disponibles aujourd'hui concernent essentiellement la conversion entre les formats stéréophonique et multicanal. Partant de ce constat, on distingue deux grandes familles de techniques de conversion de format : les techniques de «**upmix**» et les techniques de «**downmix**», selon que le système de restitution comprend plus (upmix) ou moins (downmix) de canaux que les données reçues.

Un exemple classique de technique, upmix est l'adaptation d'un signal stéréophonique pour une restitution sur un système home-cinema 5.1 (upmix 2-vers-5). Les techniques upmix permettent ainsi à l'auditeur de profiter pleinement de son système de reproduction, étant donné que la majorité des contenus proposés (CD audio, radio ou télévision) restent stéréophoniques. De plus, elles présentent l'avantage d'être «non-intrusives» puisqu'elles se situent généralement juste avant la restitution et ne modifient en rien la chaîne de transmission. Ces techniques de upmix offrent un son «optimisé» sans augmenter les débits de transmission. Il faut distinguer les techniques passives, où aucune hypothèse quant à la nature des signaux n'est émise a priori, des techniques actives qui cherchent à améliorer le traitement en analysant les propriétés des signaux. Le traitement upmix se base généralement sur une analyse dynamique des signaux d'entrées[4, 5]. Diverses hypothèses sur la composition de ces signaux sont alors formulées. Il est courant, par exemple, de chercher à extraire les composantes directes (sources individuelles

et localisables) des composantes d'ambiance (sources diffuses, lointaines et non localisables, correspondant notamment à l'effet de salle). Le traitement consiste à extraire ces composantes et à les redistribuer sur les canaux du système de reproduction: les sources directes sont redistribuées sur les trois haut-parleurs frontaux (L, C et R), tandis que les signaux d'ambiance alimentent les canaux arrière. Des signaux d'ambiance multicanaux peuvent aussi être générés par des filtres décorrélateurs qui permettent, à partir d'un unique signal, d'obtenir des signaux différents ayant le même contenu spectral mais des distributions de phase différentes, augmentant ainsi l'impression de spatialisation [6].

Les techniques de downmix impliquent une perte d'information spatiale. Le plus souvent, les algorithmes reposent sur un matriçage de type passif qui ne dépend pas de la nature des signaux d'entrée. Cependant certaines techniques sont adaptatives, par exemple dans le but de préserver l'énergie du signal et d'éviter des annulations spectrales lors de l'opération de moyennage [7]. Les techniques de downmix aujourd'hui peu utilisées en dehors du domaine du codage risquent de susciter un regain d'intérêt avec l'arrivée de nouveaux formats audio 3D.

Pour les futures technologies de conversion de format, l'universalité doit être favorisée, c'est-à-dire la faculté des algorithmes à faire abstraction du système de reproduction. Ainsi, une approche basée sur un format intermédiaire découplé du format de restitution est à étudier. Dans le même sens, développer une approche paramétrique pour décrire la configuration de haut-parleurs est primordial. L'adaptation au contenu doit aussi être prise en compte. Il est à noter que le développement des techniques de upmix souffre de l'absence de protocole de test subjectif robuste et reconnu pour évaluer leurs performances [8]. Des tests d'écoute ont cependant montré que la qualité des systèmes de upmix est inégale en regard du contenu [8].

Compression des signaux audio multicanal

Depuis l'apparition des systèmes de transmission de signaux multicanaux par matriçage comme le Dolby Surround dans les années 80 et plus récemment les Dolby Pro Logic [9], les techniques de codage du son multicanal se sont largement développées. Tout d'abord, les codeurs audio traditionnels ont été étendus à un nombre de canaux supérieur à deux (normes MPEG mp3 et AAC [10]). En parallèle, des solutions propriétaires alternatives sont apparues, telles que DTS et Dolby Digital (ou AC-3) pour le format 5.1 dans les applications grand public (par exemple pour le DVD à des débits de 384 kbit/s pour le Dolby Digital et 1.4 Mbit/s pour le DTS). Enfin, des techniques plus évoluées et complètement dédiées à l'audio 3D sont apparues à partir de la fin des années 90 avec le **codage paramétrique** permettant de proposer des schémas de codage bas débit pour des flux stéréophoniques et multicanal de type 5.1.

Ces nouvelles technologies ont été proposées récemment sous le terme de **codage audio spatial**. Dans le codeur audio spatial, un downmix (composé habituellement de un ou deux canaux) est construit à partir du signal audio original, puis codé par un codeur traditionnel (mp3, AAC, HE-AAC), et enfin transmis parallèlement avec des paramètres d'information spatiale. Cette nouvelle approche du codage audio multicanal permet de transporter ces signaux à de très faibles débits. Le procédé Binaural Cue Coding (BCC) est certainement l'un des premiers modèles de codage multicanal paramétrique proposés [11]. Il exploite les propriétés de la perception spatiale du son par l'extraction de paramètres spatiaux liés à la localisation auditive. Ces paramètres spatiaux définissent les indices de spatialisation d'une scène sonore multicanale: ICTD pour les retards entre les canaux, ICLD pour les différences d'énergie entre canaux, ICC (Inter-Channel Coherence) pour la corrélation entre les canaux. Ces paramètres sont estimés par sous-bandes de fréquences avec une résolution temps/fréquence qui suit les propriétés de la perception. Ce principe a été appliqué à la stéréophonie au travers de la norme **Parametric Stereo** [12]. L'application de cette technique avec un codeur monophonique HE-AAC permet d'obtenir un encodage stéréophonique à partir de 24 kbit/s. Plus récemment, le groupe ISO/MPEG a normalisé un format d'encodage multicanal paramétrique dénommé MPEG Surround [13]. Ce format de codage s'appuie sur des principes identiques au BCC, avec une mise en application proche du Parametric Stereo. Associé au HE-AAC, le MPEG Surround permet d'encoder le format 5.1 à des débits allant de 48 kbit/s à 160 kbit/s avec de bonnes performances en termes de qualité. La principale évolution à venir du codage audio 3D portera certainement sur le codage des futurs formats audio 3D (notamment HOA).

Vers une nouvelle génération de contenus audio 3D

Avec les technologies multicanal peut-on vraiment parler de son 3D ? De ce qui précède il ressort qu'avec les formats multicanal la spatialisation sonore reste limitée au plan horizontal en privilégiant la zone frontale. Il convient donc de faire la distinction entre ces formats et les technologies audio 3D (au sens complet de la 3D) qui offrent la possibilité

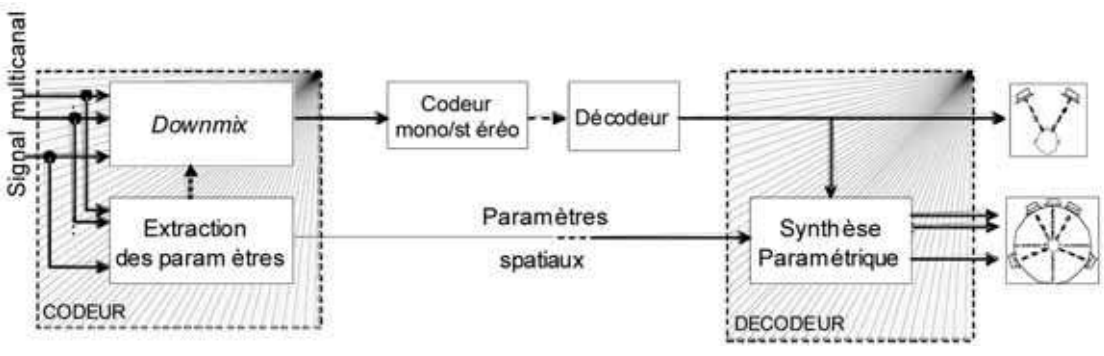


Fig. 3 : Compression des signaux audio multicanal

de restituer, simuler et contrôler des sources sonores virtuelles dans n'importe quelle direction autour de l'auditeur. Les principales technologies disponibles aujourd'hui sont: la technologie binaurale, l'holophonie (WFS) et Ambisonics (HOA). Ces dernières sont encore méconnues du grand public. Cette vision mérite d'être complétée en ajoutant qu'une autre différence entre les technologies multicanal et les technologies audio 3D (binaural, WFS, HOA) est que ces deux familles s'inscrivent dans deux philosophies distinctes. Les technologies multicanal sont des technologies issues du monde audio professionnel et sont ainsi pratiquées principalement à des fins artistiques. L'objectif n'est pas de créer une copie conforme d'une scène réelle, mais d'en donner une image, voire une interprétation. Au contraire les technologies audio 3D se basent sur un mode de représentation qui se veut conforme à la réalité acoustique de la scène de référence. Elles visent à recréer les sensations auditives naturelles, c'est-à-dire à fournir au système auditif des indices de localisation auditive qui soient conformes et fidèles à la scène de référence. En quelque sorte le son 3D est né de nouveaux besoins, liés notamment au monde de la réalité virtuelle: l'exigence de l'imitation la plus parfaite de la réalité (acoustique et/ou psychoacoustique). Ce souci est présent dès l'encodage avec des systèmes qui cherchent à exploiter de façon optimale les informations spatiales captées. Les nouvelles technologies posent la question d'une nouvelle génération de contenus. Nous allons voir comment les technologies binaurales, WFS et HOA préparent cette évolution.

Binaural

Au quotidien, nous localisons les sources sonores qui nous entourent à partir des deux signaux acoustiques captés par les tympans, qui portent en eux toutes les informations nécessaires à une perception auditive dans les trois dimensions. De ce constat sont nées les technologies dites «binaurales» : leur but n'est pas de créer un champ acoustique conforme sur une zone étendue de l'espace, mais plutôt de reproduire ou de sculpter les signaux à présenter au niveau des tympans. La restitution sur casque est idéale pour un contrôle fin de ces signaux, mais l'écoute sur deux haut-parleurs est possible ; il faut pour cela éliminer les trajets acoustiques croisés entre chaque haut-parleur et l'oreille contralatérale, celle à laquelle le signal n'est pas destiné (*cross-talk cancellation*) [14, 15].

Pour la prise de son, il suffit de placer un microphone miniature à l'entrée de chaque conduit auditif d'un individu, ou bien d'une tête artificielle. En synthèse binaurale, il s'agit de créer de toute pièce une scène sonore spatialisée à partir de sons monophoniques. Les signaux binauraux sont obtenus par un filtrage reproduisant tous les effets subis par une onde acoustique entre la source et les tympans. Ces phénomènes présentent une forte dépendance directionnelle : c'est un encodage naturel qui confère aux signaux binauraux tous les indices de localisation (ITD, ILD, modifications du timbre). Les filtres qui englobent ces phénomènes sont appelés fonctions de transfert relatives à la tête (*HRTF ou Head-Related Transfer Function*). Ils peuvent être mesurés en chambre anéchoïque (Fig. 4), ou bien calculés par des méthodes numériques, BEM ou FEM, à partir d'un maillage 3D de la morphologie de l'auditeur [16]. La synthèse binaurale est très efficace, et permet d'atteindre l'illusion parfaite dans les conditions du laboratoire [17].

Cependant, des problèmes interviennent dès que l'on s'en écarte. Si la calibration des transducteurs n'est pas rigoureuse [18], ou bien si les HRTF ne sont pas celles de l'auditeur [19], des artefacts apparaissent: perception intracrânienne, distorsion de la perception en élévation, confusion entre avant et arrière. L'utilisation de HRTF individuelles apparaît donc nécessaire, car elles portent «l'empreinte acoustique» de la morphologie de l'auditeur, à laquelle son système auditif s'est adapté. Les techniques usuelles d'acquisition des HRTF sont malheureusement inenvisageables pour une diffusion grand public. Il reste à se doter de méthodes indirectes, par exemple en s'appuyant sur des données anthropométriques ou la perception de l'auditeur lui-même [20-23]. Une application du binaural est le downmix binaural qui consiste à adapter des contenus multicanal à une écoute au casque en simulant des haut-parleurs virtuels pour les terminaux mobiles (téléphone, PDA, etc). Pour augmenter le réalisme et la sensation d'immersion, la synthèse binaurale peut être implémentée sous sa forme dynamique : un head-tracker capte les mouvements de la tête pour mettre à jour les filtres en temps réel. Les sources paraissent alors rester fixes dans une scène que l'auditeur explore en tournant la tête. Les indices de localisation supplémentaires apportés [24] permettent en outre de réduire les artefacts liés aux HRTF non-individuelles.

Fig. 4 : Mesure de HRTF en chambre anéchoïque [Pernaux]

On retiendra des technologies binaurales qu'elles présentent l'avantage de la légèreté ; en prise de son et en synthèse, les moyens matériels et logiciels nécessaires sont peu coûteux, et faciles à déployer. Leur faiblesse réside dans ce qu'il reste d'humain au cœur de la technologie ; les progrès à venir s'appuieront donc sur une compréhension toujours plus fine des mécanismes psychologiques et physiologiques de la localisation auditive.

Holophonie et WFS

La technologie holophonique, dont le concept WFS (WaveField Synthesis) est un exemple de mise en œuvre, est l'équivalent acoustique du procédé holographique [19]. Fondamentalement, elle se base sur le principe de Huygens : lorsqu'une onde acoustique se propage, chaque front d'onde peut être vu comme une distribution de sources secondaires émettant des ondelettes dont la superposition reconstruit l'onde primaire (Fig. 5). À la prise de son, un système holophonique utilise un réseau de microphones pour capter l'amplitude et la phase de l'onde acoustique sur une surface. À la restitution, les microphones sont remplacés par des haut-parleurs qui sont alimentés par les signaux microphoniques et reconstruisent ainsi l'onde acoustique originale. Le procédé holophonique est très simple dans son principe et n'implique pas de traitement, l'essentiel du travail de reconstruction des ondes sonores étant effectué par des processus «naturels» de propagation acoustique. La seule difficulté réside dans la mise en œuvre de réseaux comportant un grand nombre de transducteurs. Cependant, cette contrainte peut souvent être relâchée si on se représente le réseau de haut-parleurs comme une fenêtre ouverte sur la scène sonore. Plus cette fenêtre est grande, plus l'auditeur est immergé dans la scène sonore. Ainsi plusieurs déclinaisons du système holophonique sont disponibles selon le contenu de la scène sonore : rendu 3D (réseau de transducteurs entourant complètement l'auditeur) pour restituer des sources dans tout l'espace, rendu 2D horizontal (réseau restreint au plan horizontal), rendu frontal (rampe de haut-parleurs devant l'auditeur). La technologie holophonique a longtemps été un concept théorique qui n'avait jamais été mis en œuvre. A la fin des années 90, l'Université Technologique de Delft a proposé le premier système holophonique avec le concept de WFS [25].

Le principal atout de l'holophonie est la taille de la zone d'écoute qui n'est plus limitée à un point (*sweet point*) comme pour les technologies multicanal, mais s'étend à tout l'espace compris à l'intérieur des haut-parleurs, ce qui permet à l'auditeur de se déplacer tout en conservant une perception stable et naturelle de la scène audio 3D. L'holophonie est typiquement dédiée à un rendu multi-auditeurs, comme les salles de cinéma [26]. Récemment, la technologie WFS a exploré la nouvelle technologie des haut-parleurs plans de type MAP (Multi Actuator Panel) inspiré des DML (Distributed Mode Loudspeaker) [27] qui offrent une opportunité de mettre en œuvre la WFS en tapissant les murs d'une salle par ces panneaux vibrants. Des traitements d'égalisation multicanale sont alors appliqués afin d'optimiser le rendu spatialisé [28]. Dans les perspectives, une nouvelle approche consistant à coupler le procédé WFS avec des techniques de contrôle actif semble prometteuse [29].

HOA (Higher Order Ambisonics)

Dans leur principe et leur mise en œuvre, les technologies HOA et WFS sont proches. Inventée au début des années 70 par M. Gerzon [30], la technologie Ambisonics a été bâtie autour d'un **format de représentation intermédiaire** de la scène sonore : le Format-B. Le principe **d'encodage spatial** repose sur une captation coïncidente par une figure omni (composante W) et trois bidirectives (X, Y, Z) (Fig. 6-b). Le format de représentation qui en résulte n'est assujetti à aucun dispositif de restitution particulier. La technologie Ambisonics généralise ainsi les systèmes coïncidents de captation multicanal puisqu'elle peut simuler n'importe lequel d'entre eux, par combinaison des directivités d'encodage associées.
Le **décodage spatial** se définit comme l'opération inverse de l'encodage spatial dans la mesure où il permet de reconstituer au centre d'un dispositif de haut-parleurs la réalité acoustique captée et représentée par le Format-B. Il garantit donc la fidélité de l'organisation spatiale de la scène encodée. Au final, chaque haut-parleur restitue une portion d'espace qui aurait été virtuellement captée par un microphone hyper-cardioïde pointant dans la même direction (Fig. 6-c). En dépit de ses propriétés avantageuses, Ambisonics n'a pas connu le déploiement escompté pour des raisons conjoncturelles voire politiques [31]. Il faut aussi reconnaître que l'approche Ambisonics souffre d'une séparation spatiale limitée. De ce fait, la reproduction fidèle d'un front d'onde est réservée à une position d'écoute centrée (*sweet spot*) et aux basses fréquences (jusqu'à environ 600 Hz). Outre ce défaut de robustesse, les images sonores floues (Fig. 6-d) et l'enveloppement limité font que beaucoup de preneurs de son préfèrent à Ambisonics des approches non-coïncidentes.

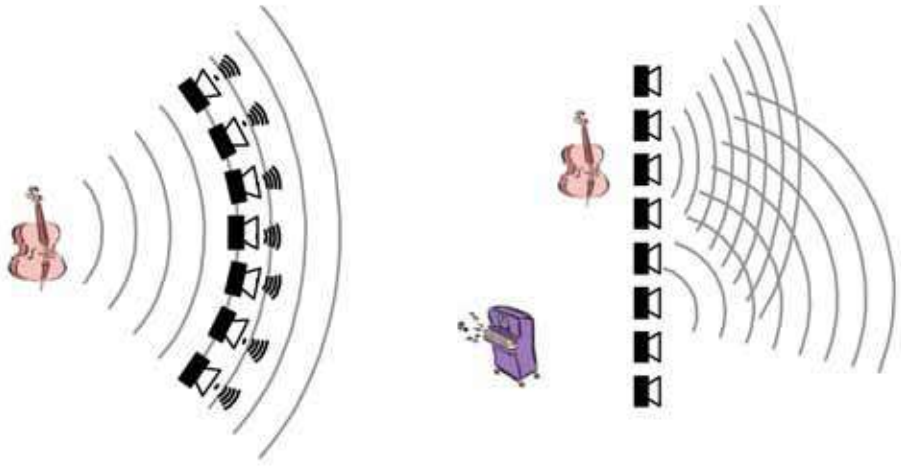


Fig. 5 : Illustration de la synthèse de fronts d'onde

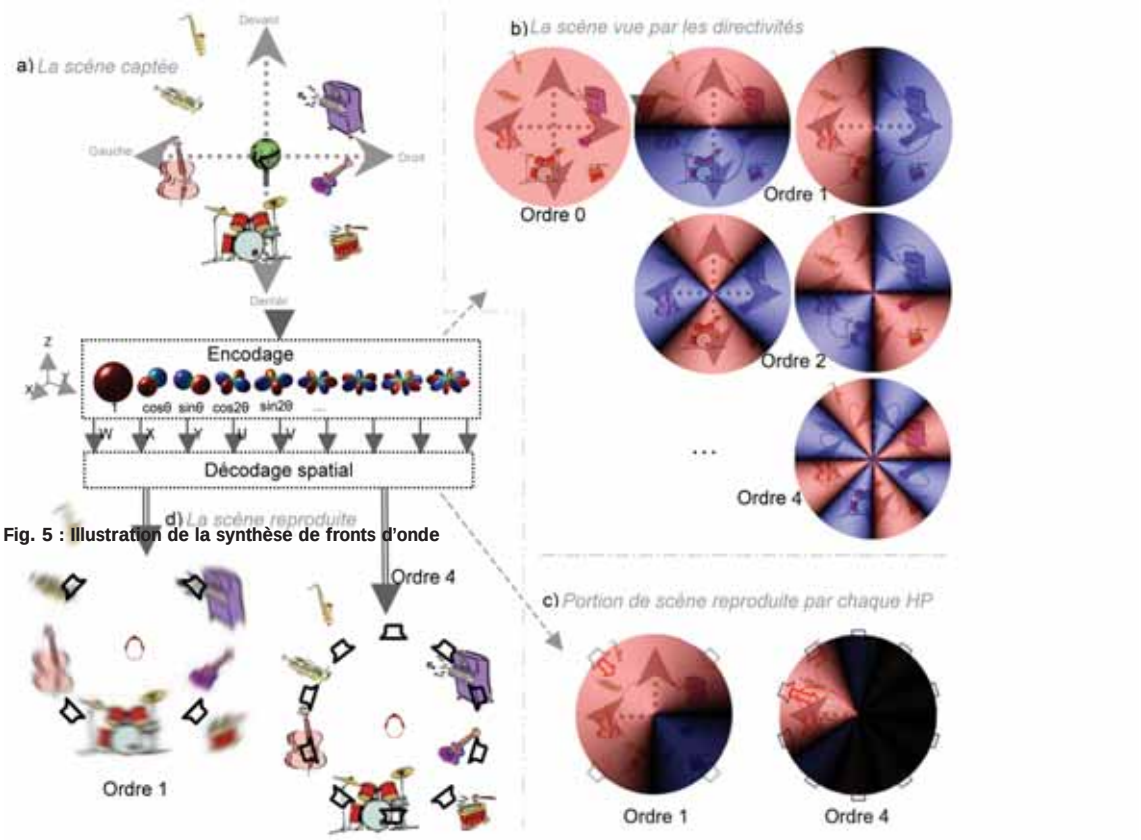


Fig. 5 : Illustration de la synthèse de fronts d'onde

Fig. 6 : Schématisation de la «chaîne» Ambisonics/HOA et illustration des principes d'encodage et de décodage, avec restriction au plan horizontal.

Vers la fin des années 90, Ambisonics a été revisitée comme une décomposition en **harmoniques sphériques** [32]. Alors que le Format-B est limité à l'ordre 1, la technologie HOA est née de l'extension de cette représentation à un ordre plus élevé. **L'encodage HOA** introduit des directivités supplémentaires (harmoniques sphériques) permettant un découpage angulaire de la scène sonore de plus en plus fin (Fig. 6-b).

En d'autres termes, il réalise une Transformée de Fourier (circulaire ou sphérique) du panorama sonore, et les signaux du format HOA constituent un **spectre spatial** dont la largeur de bande se définit par la fréquence angulaire la plus élevée. Un spectre plus riche permet une séparation plus performante et, par une mise à contribution plus sélective des haut-parleurs (Fig. 6-c), procure des images sonores plus nettes et plus stables (Fig. 6-d). La reconstruction acoustique est alors obtenue sur une zone, à la fois fréquentielle et spatiale, plus large. En pratique, l'encodage HOA est réalisé par une sphère de microphones (Fig. 6-a) qui effectue un échantillonnage spatial de la pression acoustique en vue de la projeter sur les harmoniques sphériques [33,34]. Des outils d'encodage artificiel sont également disponibles.

Le format HOA se distingue par ses propriétés d'**universalité** (restitution sur une large diversité de systèmes : casque, dispositifs 2D ou 3D de 2 à N haut-parleurs), de **scalabilité** spatiale offrant une adaptation au débit et/ou ressources disponibles (les premières composantes suffisent à décrire grossièrement la scène audio 3D, les composantes d'ordre supérieur ne font que compléter et préciser l'information spatiale), et de **flexibilité** (restitution interactive afin de manipuler la scène sonore).

Quel avenir pour l'audio 3D ?

L'avenir de l'audio 3D est riche de promesses. Les nouvelles technologies audio 3D permettent de revisiter les concepts de la spatialisation sonore qui avaient été proposés par les formats multicanal. Il ne s'agit pas d'approches concurrentes, mais plutôt complémentaires à la fois dans leurs principes fondamentaux et dans leur utilisation. On retiendra que la plupart des technologies présentent un degré avancé de maturité, du moins en ce qui concerne la captation (tête artificielle et microphone HOA) et la restitution d'une scène audio 3D (systèmes multi haut-parleurs variés ou casque). Dans le futur, on peut prévoir que les systèmes de captation et restitution audio 3D vont bénéficier des progrès des nouvelles générations de transducteurs, notamment en termes de miniaturisation. Il faut noter aussi que l'audio 3D n'est pas forcément une technologie complexe, comme l'illustrent les technologies binaurales. Un résultat tout aussi important est l'émergence, avec le format HOA, d'un format audio 3D universel, proposant une représentation à la fois flexible et scalable d'une scène audio 3D. Pour ces raisons, on peut s'attendre à ce que le son 3D fasse ses premiers

pas dans les applications grand public dans les prochaines années. Pour une maturité complète, les technologies audio 3D doivent cependant se doter d'outils de conversion de formats et de compression adaptés au transport des futurs contenus audio 3D. Une solution pressentie, dont le format HOA est précurseur, est la convergence vers un format universel intégrant des fonctionnalités de flexibilité du décodage spatial et de compression. Cette étape ne sera sans doute pas franchie avant dix ans.
L'évolution ultime de l'audio 3D sera peut-être de lever totalement les contraintes des systèmes de captation et de restitution en utilisant des réseaux hétérogènes de microphones et de haut-parleurs. Il s'agira par exemple d'extraire les informations spatiales à partir des signaux issus d'une distribution non contrainte de capteurs. Des outils d'analyse de scène sonore devront être développés. À la restitution le problème consiste à compenser ou à prendre en compte les erreurs de positionnement des haut-parleurs.

Mais pour tenir toutes ses promesses, l'audio 3D doit lever un verrou majeur qui consiste à convaincre le monde audio professionnel qui reste très attaché aux formats multicanal. Pour cela, il faut d'abord développer des outils de production et de postproduction dédiés aux nouveaux formats audio 3D afin que les ingénieurs du son s'approprient ces nouvelles technologies. L'intégration des technologies audio 3D dans les normes audio est aussi une des clefs. Des études d'usage auprès des créateurs de contenu sont également à mener pour évaluer leur utilisation des nouvelles technologies audio 3D. Les contenus auto-produits (comme des enregistrements binauraux réalisés à partir d'un téléphone mobile) sont sans doute une opportunité pour favoriser la diffusion des technologies audio 3D en suscitant l'engouement du grand public.

Références bibliographiques

[1] Theile, G. Multichannel natural music recoding based on psychoacoustic principles. 2001.

[2] Blumlein, A.D., Improvements in and relating to sound transmission, sound recording and sound reproducing systems. 1931: U.K.

[3] Leakey, D.M., Some measurements on the effect of interchannel intensity and time difference in two channel sound systems. J. Acous. Soc. Am., 1959. 31: p. 977–986.

['] Avendano, C. and J.-M. Jot, A frequency-domain approach to multichannel upmix. J. Audio Eng. Soc., 2004. 52(7-8): p. 740-749.

[5] Irwan, R. and R.M. Aarts, Two-to-five channel sound processing. J. Audio Eng. Soc., 2002. 50(11): p. 914-926.

[6] Kendall, G.S., The decorrelation of audio signals and its impact on spatial imagery. Computer Music J., 1995. 49: p. 71-87.

[7] Faller, C., Parametric coding of spatial audio. 2004, Ecole Polytechnique Fédérale de Lausanne.

[8] Chétry, N., et al. A discussion about subjective methods for evaluating blind upmix algorithms. in 31st International Conference of the Audio Eng. Soc. 2007.

[9] Dressler, R., Dolby surround Pro Logic decoder principles of operation. 2000, Dolby Laboratories.

[10] Bosi, M. and R.E. Goldberg, Introduction to Digital Audio Coding and Standards. 2002, Dordrecht: Kluwer Academic Publishers.

[11] Faller, C. and F. Baumgarte. Binaural cue coding: a novel and efficient representation of spatial audio. in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP'02). 2002. Orlando, USA.

[12] Breebaart, J., et al., Parametric Coding of Stereo Audio. EURASIP Journal on Applied Signal Processing, 2005. 9: p. 1305-1322.

[13] Breebaart, J., et al. MPEG spatial audio coding / MPEG surround: overview and current status. in 119th Convention of Audio Eng. Soc. 2005. New York, USA.

[14] Gardner, M.B., 3D audio using loudspeakers. 1998: Kluwer Academic Publishers.

[15] Kirkeby, O., P.A. Nelson, and H. Hamada, The 'stereo dipole' : a virtual source imaging system using two closely spaced loudspeakers. J. Audio Eng. Soc., 1998. 46(5): p. 387-395.

[16] Kahana, Y., Numerical modelling of the head-related transfer function. 2000, University of Southampton: Southampton, UK.

[17] Martin, R.L., K.I. McAnally, and M.A. Senova, Free-field equivalent localization of virtual audio. J. Audio Eng. Soc., 2001. 49(1/2): p. 14-22.

[18] Pralong, D. and S. Carlile, The role of individualized headphone calibration for the generation of high fidelity virtual auditory space. J. Acous. Soc. Am., 1996. 100(6): p. 3785-3793.

[19] Nicol, R., Restitution sonore spatialisée sur une zone étendue: Application à la téléprésence. 1999, Université du Maine: Le Mans.

[20] Zotkin, D.N., et al. HRTF personalization using anthropometric measurements. in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA 2003. 2003. New Patlz, NY, USA.

[21] Middlebrooks, J.C., Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency. J. Acous. Soc. Am., 1999. 106(3): p. 1493-1510.

[22] Middlebrooks, J.C., E.A. MacPherson, and Z.A. Onsan, Psychophysical customization of directional transfer functions for virtual sound localization. J. Acous. Soc. Am., 2000. 108(6): p. 3088-3091.

[23] Martens, W.L. Rapid psychophysical calibration using bisection scaling for individualized control of source elevation in auditory display. in Proc. Int. Conf. on Auditory Display, ICAD 2002. 2002. Kyoto, Japan.

[24] Wightman, F.L. and D.J. Kistler, Resolution of front-back ambiguity in spatial hearing by listener and source movement. J. Acoust. Soc. Am., 1999. 105(5): p. 2841-2853.

[25] Berkhout, A.J., A holographic approach to acoustic control. J. Audio Eng. Soc., 1988. 36(12): p. 977-995.

[26] Sporer, T. Wave Field Synthesis – Generation and reproduction of natural sound environments. in DAFX'04. 2004. Naples.

[27] http://www.nxtsound.com/.

[28] Corteel, E., Adaptation de la Wave Field Synthesis aux conditions réelles. 2004, Université de Paris 6.

[29] Gauthier, P.A., Synthèse de champs sonores adaptative. 2007, Universite de Sherbrooke.

[30] Gerzon, M.A., Periphony: With-Height Sound Reproduction. J. Audio Eng. Soc., 1973. 21(1): p. 2-10.

[31] Elen, R., Whatever happened to Ambisonics ? (trad. fr. : http://fgouget.free.fr/ambisonic/Ambisonic_AM91-fr.shtml), in Audio Media Magazine. 1991.

[32] Daniel, J., Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia. 2000, Université Pierre et Marie Curie (Paris VI): Paris.

[33] Moreau, S., Étude et réalisation d'outils avancés d'encodage spatial pour la technique de spatialisation sonore Higher Order Ambisonics : microphone 3D et contrôle de distance. 2006, Université du Maine: Le Mans.

[34] Elko, G. and J. Meyer, Electroacoustic Systems for 3-D Audio - A report from the Pittsburg meeting. Echoes, 2002.