Dispositif d'Aide à la Spatialisation basé sur des critères perceptifs

Laurent Simon

Stage de Recherche
Master Recherche ATIAM
Université Pierre et Marie Curie, IRCAM
Juin 2006

Coencadré par Olivier Delerue et Olivier Warusfel

Table des matières

Remerciements

Il ne s'agit là que d'un rapport de stage, mais je tenais tout de même à effectuer quelques remerciements. Merci tout d'abord aux Oliviers (non non, pas les arbres, les gens) qui m'ont encadré et appris des très bonnes méthodes que j'espère avoir su appliquer au cours de ce stage.

Merci à Olivier Delerue pour m'avoir également permis de toucher à tout plein de choses (la souris, le clavier, trois ordinateurs dont un qui est resté en vie, Google Earth, Larousse et Hachette, et accessoirement des outils de travail puissants...) et avoir eu la bonté de faire semblant de rire à certains de mes jeux de mots vaseux. Merci plus généralement à l'équipe d'Acoustique des Salles, aux ATIAM et plus particulièrement à ceux qui ont mangé avec moi le midi pour m'avoir supporté, et aux pièces rapportées qui bien que ne faisant partie d'aucune de ces deux catégories, ont subit mes blagues quand même (n'est-ce pas, Julien, Guillaume et Antoine ?). Merci aussi au Bo Bun qui était drôlement bon et qui fait sa rentrée le jour du rendu de ce rapport, à la pizza savoyarde qui me collait à ma chaise et aux sandwichs divers qui m'ont permis de survivre dans la jungle grouillante du quartier de Beaubourg.

Chapitre 1

Introduction

L'objectif de ce stage est de contribuer à un système d'aide à la spatialisation du son qui propose la représentation de critères de qualité d'une scène sonore afin de guider l'utilisateur dans ses choix. Ces critères de qualité sont calculés à partir de facteurs perceptifs et d'indices acoustiques de sorte à traduire objectivement la qualité perçue. Le principe de cet outil est de représenter, en arrière plan d'une interface bidimensionnelle de visualisation et de manipulation d'une scène sonore, l'évolution d'un critère de qualité en fonction de la position des sources et / ou de l'avatar. Le système permet ainsi à son utilisateur de prédire quel sera l'impact d'une modification de la scène sonore au regard du critère de qualité sélectionné.

Nous dresserons dans un premier temps un panorama d'outils de contrôle de la spatialisation existants en nous intéressant à la part que ceux-ci réservent à l'aide au mixage et à la spatialisation. Nous nous intéresserons également aux techniques de spatialisation et aux systèmes de restitution du son correspondants afin d'étudier leur influence sur les critères de qualité mis en œuvre, critères de qualité basés sur un ou plusieurs attributs perceptifs pouvant être déduits d'une analyse objective de la scène auditive.

Nous présenterons dans un deuxième temps les facteurs perceptifs qui nous semblent intéressants pour l'élaboration de critères de qualité de scènes sonores : le masquage, la notion d'image fantôme et la perception d'artefacts lors de la spatialisation de sources démixées. Nous présenterons également les travaux associés sur lesquels nous nous sommes basés.

Nous décrirons l'implémentation de ces travaux sous forme d'une boîte à outil Matlab ainsi que sous forme de classes Java intégrées au sein de l'outil d'aide au mixage existant dans l'application ListenSpace.

Finalement nous présenterons les travaux réalisés dans le cadre de la validation perceptive de cette mise en œuvre puis nous terminerons cette étude en dégageant les orientations futures et améliorations possibles de ces recherches.

Chapitre 2

Etat de l'Art

Dans cette partie, nous réalisons tout d'abord un état de l'Art sur différents systèmes de restitution audio : le binaural, le système le plus proche de notre écoute, la stéréo, système le plus répandu et sur lequel repose le répertoire musical le plus large, et enfin le 5.1. Nous nous attardons ensuite sur les outils d'aide à la spatialisation existants : des outils donnant des informations sur la qualité de la scène sonore, des outils simplifiant les paramètres de contrôle d'une scène sonore spatialisée ou des outils appliquant des contraintes sur les différentes sources afin de conserver une cohérence du mixage. Enfin, nous faisons un bilan des publications concernant les facteurs perceptifs qui nous intéressent, facteurs perceptifs qui seront traités plus en profondeur dans la partie 4.

L'attitude du grand public face à un enregistrement musical a longtemps été passive : les seules actions que l'auditeur pouvait réaliser étaient la mise en marche ou l'arrêt de la lecture de l'enregistrement ainsi que le changement du niveau global de la lecture ou l'égalisation du système de diffusion. Il était bien sûr possible de déplacer une enceinte, de lui appliquer un retard ou un gain particulier, mais de telles modifications n'étaient pas évidentes.

Avec l'apparition du codage par le contenu (norme MPEG 4), l'auditeur peut désormais interagir avec un contenu musical. Le codage par le contenu consiste à coder séparément les différentes sources d'un enregistrement et à joindre une piste de description de la scène sonore, souvent appelée piste de metadonnées consignant des informations relatives au mixage de ces différentes sources : positions dans l'espace, description de l'acoustique de l'espace, effets, mouvements, ...

Le codage séparé des sources permet, moyennant des outils adéquats, de déplacer ces sources dans l'espace, de modifier les effets appliqués et les caractéristiques de l'espace. C'est alors en bout de chaîne que le choix du

système de restitution et le traitement qui en découle sont effectués, ce qui rend ce format indépendant du système de restitution et autorise un vaste champ d'intéraction pour l'utilisateur final. Ces idées rentrent dans le cadre du projet européen SemanticHIFI qui vise à concevoir une chaîne Hi-Fi permettant, entre autres, un classement personnalisé des morceaux de musique et l'intéraction avec le contenu, notamment au niveau de la spatialisation [1]. La question qui se pose alors est : *Comment produire des outils permettant à des utilisateurs non experts d'exploiter intelligemment / efficacement ces nouvelles fonctionnalités ?*.

Outre la description des principaux systèmes de diffusion, nous nous intéresserons aux outils d'aide à la spatialisation existants.

2.1 Systèmes de restitution

L'enjeu du dispositif étudié étant de prédire et visualiser la qualité acoustique de la scène perçue par l'auditeur, il convient de choisir le point d'observation de cette qualité ainsi que le contexte d'écoute. Il est par conséquent proposé d'effectuer l'expertise en se plaçant aux oreilles de l'auditeur.

Dans le cas d'une restitution en mode binaural, le système de restitution est supposé transparent : on ne tient compte que des sources virtuelles et de leur position désirée. Dans le cas de la stéréo ou du 5.1, on doit en revanche appliquer la loi de panpot sur les sources virtuelles pour connaître le signal délivré par chaque enceinte puis une synthèse binaurale prenant ces deux enceintes comme sources virtuelles situées à +30°et -30°(-110°, -30°, 0°, +30°et +110°dans le cas du 5.1) afin de prédire le champ acoustique aux oreilles du sujet.

2.1.1 Binaural

Le binaural est un système de prise de son / synthèse / restitution utilisant deux canaux audios. Lors de la restitution, chaque canal est envoyé respectivement dans une oreille de l'auditeur par le biais d'un casque audio. Le but de ce système est de reproduire tous les indices de perception des sons dans l'espace, aussi bien d'un point de vue angulaire que de celui de la distance.

On recense deux indices importants pour la localisation angulaire d'une source :

- La différence interaurale de temps (ITD ou Interaural Time Difference), due à la différence de longueur de trajet de la source à chacune des deux oreilles
- La différence interaurale d'intensité (IID pour Interaural Intensity Difference ou ILD pour Interaural Level Difference) due aux diffractions, réflexions et à l'absorption des ondes acoustiques par le corps de l'individu. Elle est dépendante de la fréquence.

Cette dichotomie provient de la théorie duplex, élaborée par Lord Rayleigh en 1907 [27] qui visait à une description simple des mécanismes de localisation en azimut.

La combinaison de ces deux indices forme ce que l'on appelle généralement les HRTF (Head Related Transfer Functions, dans le domaine fréquentiel) ou HRIR (Head Related Impulse Responses, dans le domaine temporel).

Il s'agit des fonctions de transfert caractérisant le canal acoustique d'une source jusqu'à chacune des deux oreilles d'un sujet pour une position spatiale de la source donnée. Elles sont indépendantes de la distance de la source (à un gain près) dès que la source est éloignée d'au moins 1 mètre (en deçà de cette distance, un phénomène de champ proche apparaît). Elles peuvent être mesurées sur un sujet ou un mannequin ou encore synthétisées [3].

Pour réaliser un enregistrement binaural, il est possible

- soit d'enregistrer les signaux de pression acoustique incidente au niveau des oreilles d'un individu ou d'une tête réelle (méthode de prise de son)
- soit de mesurer les HRTFs d'un individu puis de synthétiser les signaux binauraux à partir de ces mesures
- soit de modéliser les HRTFs à l'aide de modèles physiques (morphologie de la tête) pour stimulerles fonctions de transfert binaurales
- soit de modéliser les HRTFs à l'aide de modèles analytiques (dichotomie ITD/IID puis modélisation de l'IID et de l'ITD séparément) puis synthèse des signaux binauraux

2.1.2 Stéréo

Ce système utilise généralement deux enceintes disposées à deux des sommets d'un triangle équilatéral, le dernier étant occupé par la tête du sujet (voir fig. 2.1). Les deux enceintes sont ainsi positionnées aux azimuts +30 et -30 (voir fig. 2.2). Pour spatialiser une source audio sur un tel système, on envoie le signal sur chacune des enceintes avec un niveau et/ou un retard différent.

On appelle loi de panpot l'ensemble de ces fonctions $[\Delta_{left}t, \Delta_{left}L] = f(\varphi)$ et $[\Delta_{right}t, \Delta_{right}L] = g(\varphi)$, où $\Delta_{left}t$ et $\Delta_{left}L$ sont respectivement le retard et l'atténuation en niveau de l'enceinte gauche ($\Delta_{right}t$ et $\Delta_{right}L$ pour l'enceinte droite).

Le rayonnement différent des deux sources réelles provoque des différences de phase et d'amplitude aux oreilles du sujet. Le cerveau recompose ces différences d'amplitude et/ou de phase comme des indices interauraux de spatialisation et en déduit la localisation de la source fantôme induite. Les différences interaurales que l'on peut alors mesurer ne correspondent pas nécessairement à celles d'une source réelle, en particulier au regard de la dépendance fréquentielle des indices. La source diffusée par le système stéréo n'en a pas moins une position spatiale définie. Différentes études ont été réalisées sur la position perçue d'une source en fonction des différences

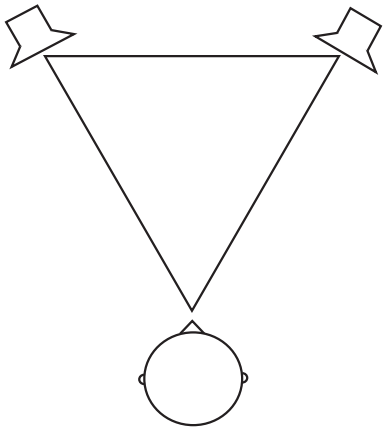

FIG. 2.1 – Configuration stéréo : les deux enceintes et l'auditeur se trouvent aux sommets d'un triangle équilatéral

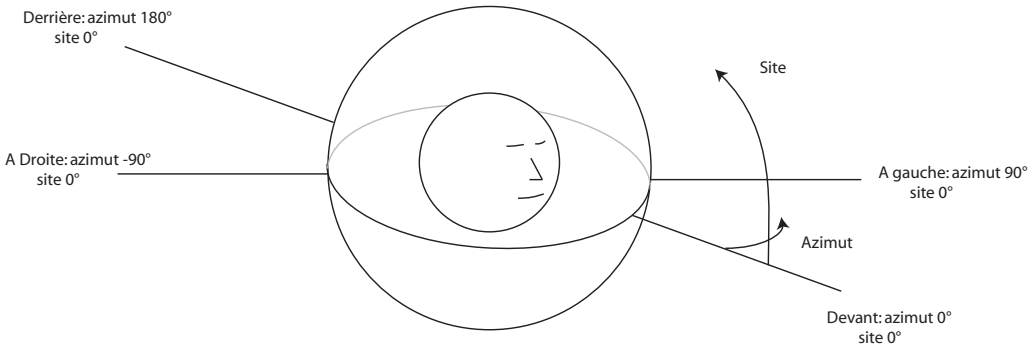

FIG. 2.2 – Le repère des angles dans l'espace. L'azimut représente l'angle d'une source dans le plan horizontal. Le site (elevation en anglais) représente l'angle vertical.

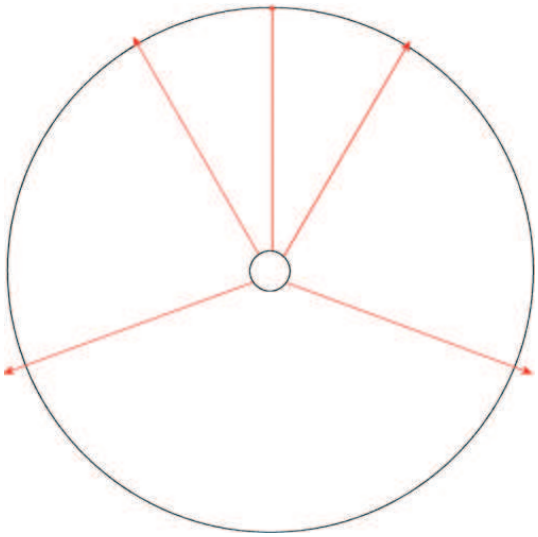

FIG. 2.3 – Angles de disposition des haut-parleurs pour un système 5.1 respectant la norme ITU-R BS.775

interaurales de phase et d'intensité [2].

2.1.3 multicannal 5.1

La stéréo ne permettant de créer des images fantôme que dans le secteur angulaire frontal situé entre les deux enceintes[1] et la quadriphonie ayant disparu faute de consensus sur le format à adopter, une alternative est apparue dans les cinémas et chez les particuliers : le système 5.1. Il est composé de cinq enceintes et un caisson de basse, placés selon la norme ITU-R BS.775 (voir fig. 2.3). Notons que dans les cinémas, cette norme est rarement respectée pour des raisons de largeur de sweet spot. Selon cette norme, les haut-parleurs Droit, Centre, Gauche, Arrière Gauche et Arrière Droit doivent être disposés respectivement à -30°, 0°, 30°, 110°et 250°. Le problème de l'intégration d'un tel système dans notre outil de représentation est le même que pour la stéréo, nous ne détaillerons donc pas plus ce sujet.

2.1.4 Autres systèmes

Nous citons également à titre informatif deux systèmes de spatialisation (et de diffusion) particuliers : la Wave Field Synthesis et le High Order Ambisonic. Ces deux systèmes visent à reproduire un champ sonore non pas

[1] à l'exception de diffusion en transaural, système assez contraignant réservé à un seul utilisateur qui ne bouge pas, ou bien d'effet de hors-phase, parfois voulus par les ingénieurs du son, mais souvent désagréabes

en un point ou une ligne (comme les systèmes présentés) mais sur une zone étendue. Au cours de ce stage, nous avons cependant privilégié le binaural et la stéréo, le premier étant le plus simple à étudier d'un point de vue perceptif (puisqu'il veut reproduire notre perception des sons dans l'espace), et le second étant le système le plus répandu.

Au cours de ce stage, nous nous intéressons à une approche perceptive de la spatialisation. Par conséquent, quel que soit le système de diffusion envisagé, nous pouvons toujours nous rapporter aux oreilles de l'auditeur par le biais d'une modélisation de l'effet de chaque enceinte en convoluant celle-ci pas la fonction de transfert binaurale associée.

2.2 Outils d'aide à la spatialisation

La mise à disposition du grand public de systèmes de spatialisation pose un problème d'expérience : les auditeurs sont rarement habitués à mixer des sources et manquent ainsi d'un certain savoir-faire. L'utilisateur risque ainsi de placer des sources à des endroits où il ne peut les entendre (masquage mutuel), de trop espacer des sources qui ne devraient pas l'être si on souhaite conserver certaines images fantômes ou de détimbrer ces sources. Il devra également manipuler des grandeurs qu'il ne comprend pas nécessairement sans pouvoir en prédire les conséquences. Une aide est alors souvent bienvenue...

Le premier outil d'aide à la spatialisation est le phasemètre, qui utilise une figure de Lissajous. Un oscilloscope représente la figure de Lissajous ayant pour entrées les canaux gauche et droit d'un mixage stéréo. Cette figure donne alors de précieuses informations sur le contenu spatial du mixage (voir fig. 2.4).

C'est cependant un outil que l'on ne trouve généralement que dans les grands studios d'enregistrement. Il n'est pas très intuitif et sert principalement à vérifier qu'il n'y ait pas de signaux hors-phase dans le mixage réalisé.

Le guidage peut consister en une mesure objective d'un facteur perceptif, comme ici l'effet de hors-phase (ce qui déplace le son sur le côté et crée une impression désagréable de son à la fois dans et hors de la tête), mais également en une meilleure lisibilité des paramètres que l'on peut faire varier. C'est là le principe du Spat_Oper, une interface de commande du SPAT ([18],[19]), un programme pour MAX/MSP développé par l'IRCAM et qui permet le traitement de sources sonores afin de les spatialiser. Il combine des outils de calcul de réverbération avec des moteurs de rendu permettant d'écouter ces sources sur différents systèmes de restitution..

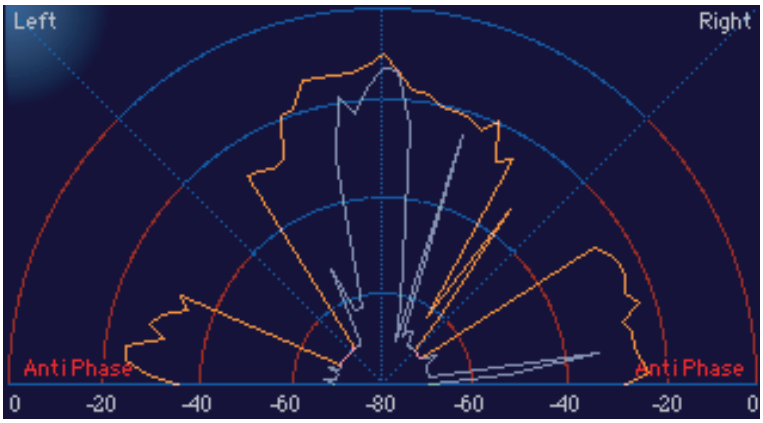

FIG. 2.4 – Un phasemètre indique la position de l'image stéréo et sa largeur. Un simple trait vertical représenterait un son monophonique. Un trait dans la directon 45 degrés horaire ou antihoraire représente respectivement un son uniquement dans le canal droit ou dans le canal gauche. Des traits dans les zones au-delà de ces angles représentent des sons hors-phase (qui donnent l'impression désagréable d'être à la fois en dehors de la tête et dedans. Il est donc préférable d'éviter ce type de situation)

2.2.1 le Spat_Oper

Le travail autour de cette interface et du programme de spatialisation qui lui est lié a consisté en une réduction du nombre de paramètres de spatialisation et est la conséquence d'une étude psycho-perceptive. Partant d'un nombre initial d'environ 100 paramètres physiques, cette étude a permis de réduire ce nombre à une dizaine de paramètres perceptifs, et alors compréhensibles par des utilisateurs non spécialisés dans la spatialisation du son.

Afin de manipuler les paramètres du SPAT, ses créateurs ont mis au point une interface nommée `Spat_Oper`. L'utilisateur contrôle les quelques paramètres à sa disposition à l'aide de curseurs similaires à ceux que l'on pourrait trouver sur une table de mixage. La différence principale avec ces dernières est que la grandeur manipulée par un curseur n'est pas de bas-niveau (niveau, filtre,...) mais un composé de plusieurs grandeurs physiques, regroupées sous une seule grandeur perceptive. Il est ainsi possible de contrôler, à l'aide de ces curseurs, la présence (de la source ou de la salle), la chaleur, la brillance, la réverbérance, l'enveloppement, ... (voir fig. 2.5). La taille des curseurs est proportionnelle à l'importance de chaque indice, telle qu'elle a été déterminée par une analyse statistique lors de la création du Spat (un indice important donnera lieu à un grand curseur, un indice moins important donnera lieu à un plus petit curseur). Cela illustre l'intérêt d'une approche perceptive : en approche traitement du signal, un paramètre peut avoir une conséquence

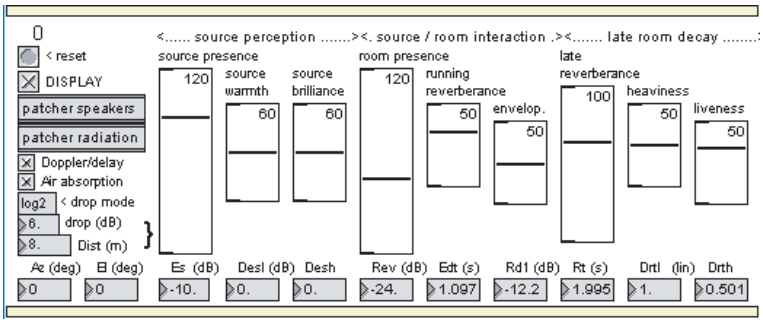

FIG. 2.5 – L'interface de `Spat_Oper`, sous MAX/MSP. Des curseurs permettent de contrôler les paramètres perceptifs de la source (présence, chaleur, ...) et ceux de la salle. Les paramètres physiques sont quand à eux contrôlés par des boîtes numériques (azimut, distance, ...)

perceptive qui diffère qualitativement et quantitativement suivant le contexte (i.e. suivant le réglage des autres paramètres). Les études psychophysiques qui ont présidé à l'élaboration de l'interface Spat_Oper permettent d'assurer une meilleure indépendance (orthogonalité) des effets manipulés et par conséquent une utilisation plus intuitive.

Cependant, cette interface ne décrit que la perception d'une source à la fois et non d'une scène constituée de plusieurs sources. Le Spat_Oper ne renseigne donc pas sur le masquage des sources, sur la localisation relative, etc...

2.2.2 MusicSpace

Le libre placement de sources sonores dans l'espace par un utilisateur non expert peut amener à des configurations aberrantes telles que des sources sonores trop espacées ou trop rapprochées. Pour y remédier, l'idée de spatialisation par contraintes a fait l'objet d'un travail de recherche et de développement [7]. L'utilisateur ne dispose alors plus librement les sources dans l'espace. Ces sources sont liées par un jeu de contraintes agissant sur les paramètres de spatialisation qui leur sont associées. Différentes idées se cachent derrière chaque contrainte. Par exemple, une contrainte angulaire qui impose que l'angle entre deux sources reste constant vise à préserver la fusion ou au contraire la lisibilité des flux sonores émanant de ces sources. Ainsi, on peut maintenir voisines les sources d'une même section instrumentale ou au contraire les maintenir écartées pour faciliter la perception d'un contrepoint. C'est le principe du logiciel MusicSpace [7] : traduire des paramètres bas niveaux (distance, azimut, vitesse de déplacement, ...) en des paramètres haut niveau (équilibre section rythmique / section mélodique d'un morceau, distances des différents instruments, ...) et permettre à l'ingénieur du son

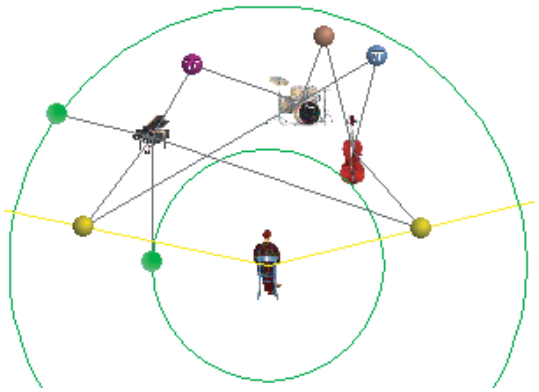

FIG. 2.6 – Interface de MusicSpace. Exemple de contraintes entre différentes sources musicales : une batterie, une contrebasse et un piano. Les contraintes sont représentées par les traits et les boules, les boules spécifiant alors le type de contrainte.

d'imposer des choix artistiques ou techniques tout en laissant à l'utilisateur une certaine liberté d'interaction (voir fig. 2.6).

Ce logiciel nécessite cependant un traitement préalable : un ingénieur du son doit créer des contraintes entre les différentes sources d'un mixage. Mais cela assure une certaine cohérence et permet d'imposer des choix artistiques des musiciens, du producteur ou de l'ingénieur du son.

2.3 L'interface de ListenSpace

ListenSpace est le logiciel de représentation de scène sonore sur lequel nous avons intégré un système d'aide au mixage. Il permet à un utilisateur de créer, modifier et visualiser un espace sonore virtuel en indiquant de nombreux paramètres. Cette scène sonore est composée d'un auditeur et d'au moins une source sonore virtuelle. L'utilisateur peut contrôler les positions et directions de ces objets, placer un repère normé pour indiquer l'échelle de la scène, placer des murs ou d'autres éléments architecturaux et découper la scène en zones d'interaction utiles dans le cas d'une utilisation en réalité virtuelle ou augmentée (voir fig. 2.7). Cette représentation se fait sans tenir compte du système de restitution. L'utilisateur représente en effet la scène sonore telle qu'il souhaite pouvoir l'entendre dans un cas idéal. Tout ce qu'il peut représenter est alors virtuel et sera simulé lors de l'utilisation d'un outil de spatialisation avec un système de restitution donné. Les enceintes réelles dont l'utilisateur dispose n'apparaissent pas sous ListenSpace.

L'objectif de ListenSpace est de prendre en charge la partie visualisation et interaction d'un environnement de réalité virtuelle. L'application s'inscrit

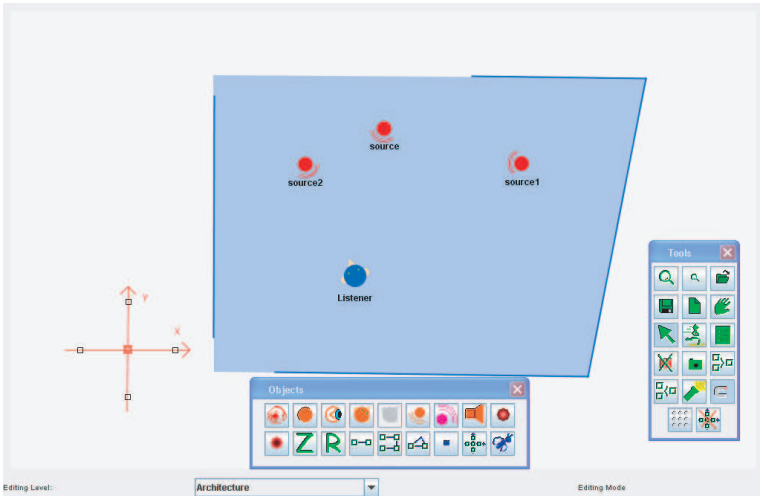

FIG. 2.7 – L'interface principale de ListenSpace. On y distingue un auditeur (le Listener) ainsi que plusieurs sources sonores (les points). L'utilisateur peut spécifier leur position et leur direction, ainsi que décrire une architecture autour de ces sources.

donc, au moyen de communications réseau, au sein d'un ensemble de composants logiciels qui gèrent chacun leur tâche (tracking, joystick, comportement, rendu sonore ou vidéo, ...). C'est pour cela qu'il peut échanger des informations avec des logiciels tiers, qui se chargent alors de lire les fichiers audios et les métadonnées qui lui sont associées, d'envoyer ces métadonnées à ListenSpace et de recevoir de l'outil de visualisation les informations concernant la scène sonore et de traiter les signaux audio pour transposer cette scène sur le système de diffusion de l'utilisateur. Pour réaliser cet échange de données, ListenSpace utilise le protocole UDP. On spécifie à ListenSpace les adresses UDP desquelles il doit recevoir des données et celles auxquelles il doit en envoyer (voir fig 2.8). Actuellement, il est principalement utilisé avec MAX/MSP et le Spat.

Comme nous l'avons dit précédemment, pour un utilisateur inexpérimenté, la création d'une scène sonore cohérente n'est pas une tâche évidente. De nombreux paramètres sont à prendre en compte, et le rendu perceptif s'éloigne parfois de celui que l'utilisateur souhaitait. Il est donc nécessaire de guider l'utilisateur lors de son mixage, et c'est par la visualisation de facteurs perceptifs que nous avons choisi de le faire.

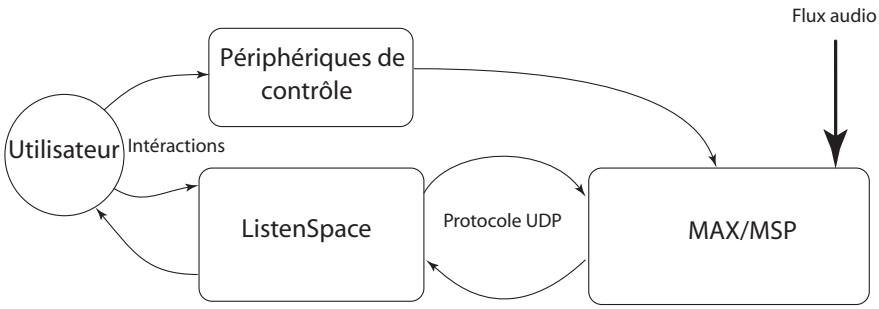

FIG. 2.8 – Schéma d'interaction entre ListenSpace et MAX/MSP via un protocole UDP. Cela permet d'utiliser le Spat comme moteur de rendu audio 3D

Chapitre 3

Principes de l'outil de spatialisation guidée

Nous présentons dans cette partie les principes généraux de l'outil d'aide à la spatialisation envisagé. La visualisation se fait à l'aide de niveaux de gris indiquant en tous points de la scène sonore si la qualité est bonne (pixels plus ou moins clairs) ou mauvaise (pixels plus ou moins foncés). Ces critères doivent pouvoir être représentés en temps réel et permettre à l'utilisateur de prévoir les conséquences de ses actions.

Dans cette partie, nous nous intéressons à la visualisation de critères de qualité dont le but est l'aide au mixage d'une scène sonore en temps réel. L'outil de visualisation sur lequel nous avons travaillé est intégré à l'interface ListenSpace, décrite dans la partie 2.3.

Nous avons choisi de guider l'utilisateur via la représentation en temps réel de facteurs perceptifs dans l'interface. Cette représentation de facteurs perceptifs tels que l'homogénéité spatiale ou le masquage spatial vise à donner à l'utilisateur de ListenSpace une information sur la qualité future de la scène sonore à l'aide d'un modèle d'évaluation d'indices perceptifs / de critères de qualité évolués ou en fonction des signaux associés aux différents éléments constitutifs de la scène, de leur organisation spatiale et du contexte d'écoute (dispositif d'écoute et modèle de spatialisation ; voir les organigrammes fig. 3.3 et 3.4). Pour donner ces informations, l'interface utilise des niveaux de gris appliqués sur le fond d'écran de l'interface, rarement utilisé par les autres outils de représentation de scènes sonores. L'utilisateur peut ainsi prévoir l'effet du placement de sources dans sa scène sonore et la modifier en conséquence (voir fig. 3.1 et 3.2 pour des exemples de représentation).

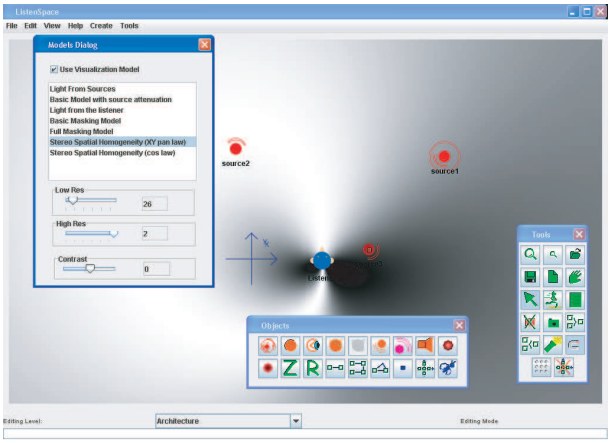

FIG. 3.1 – Représentation de l'équilibre spatial sous ListenSpace : On veut savoir où placer la source entourée (en haut à droite du Listener) de façon à ce que la scène soit équilibrée spatialement. Les zones claires sont les endroits où le placement de la source cible serait optimal. Les zones sombres indiquent que la scène sonore ne serait plus équilibrée si la source cible venait à être placée à ces endroits.

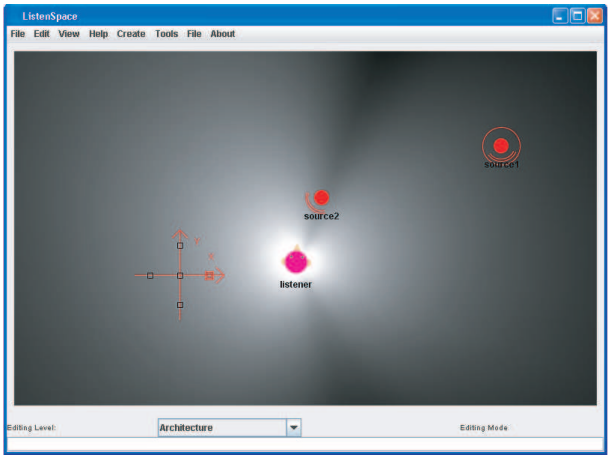

FIG. 3.2 – Représentation du masquage spatial sous ListenSpace : On veut savoir où placer la source 1 entourée (en haut à droite de l'auditeur) de façon à ce qu'elle ne soit pas masquée par la source 2.

Paramètres d'évaluation

Nous nous intéressons au calcul d'indices perceptifs / critères de qualité. Dans le domaine auquel nous nous adressons, un indice est une mesure objective d'un phénomène psychophysique. Il renseigne sur une propriété de la perception sans porter de jugement subjectif (préférence). Un exemple est fourni fig.3.2 où l'on représente l'indice de masquage d'une source en fonction d'une autre. Dans un second temps, on peut également vouloir représenter des critères de qualité davantage liés à une notion de préférence comme dans le cas de la fig.3.1 qui représente l'équilibre spatial de la scène sonore.

Dépendance spatiale

L'évaluation d'un critère de qualité permet d'aider un utilisateur par le biais d'une interface perceptive : il est possible de prédire l'évolution d'un indice en fonction du placement de sources sonores dans l'espace en un point donné, le point d'écoute de l'utilisateur. Afin de ne pas se limiter à informer l'utilisateur de la qualité de la scène sonore basée sur un certain critère, il est proposé de choisir une source cible parmi les différentes sources de la scène sonore par le biais d'un mécanisme de sélection et d'évaluer le critère de qualité pour toutes les positions possibles de cette source cible. L'utilisateur est ainsi informé non seulement de la qualité acoustique pour la position courante de l'objet sélectionné mais également de la valeur qu'aurait pris le critère si la source cible avait été placée ailleurs qu'en son point actuel (voir fig.3.1 et 3.2).

Dépendance temporelle

Afin que l'utilisateur puisse anticiper l'évolution de sa scène sonore, il est nécessaire de prédire l'évolution des indices perceptifs. L'idée de la prédiction est de calculer les indices en avance de quelques secondes et de les stocker jusqu'à ce qu'ils soient obsolètes. On applique ensuite une fenêtre de pondération sur ces indices pour les moyenner dans le temps en tenant compte de la pertinence de ces indices : l'indice calculé correspond prioritairement à ce que l'auditeur entend au moment de la représentation mais renseigne également sur l'état à venir. En pratique, cela nécessite de disposer des métadonnées à l'avance pour lire dans le futur. Par ailleurs, le choix de la fenêtre de prédiction résulte d'un compromis entre la réactivité du système (conséquence des fluctuations liées au contenu des sources) et sa lisibilité.

Dépendance au contexte

Le calcul des indices en un point d'écoute dépend donc de ce qui arrive en ce point d'écoute et ce que perçoit l'auditeur. Il faut donc tenir compte du modèle de spatialisation du son, du système de restitution et transposer

le modèle perceptif adopté sur ce système. Le modèle de spatialisation décrit comment est traitée une source sonore en fonction de la scène sonore et du système de restitution : simple modèle d'atténuation en distance et panpot d'intensité ou bien filtrage par HRTF, prise en compte des caractéristiques d'une salle virtuelle, prémières réflexions, champ diffus, ...

Typiquement, le paramètre perceptif choisi sera noté en niveau de gris en tous les points de la scène sonore, un gris foncé indiquant que la scène sonore ne sera pas satisfaisante selon ce critère si la source cible est placée en ce point tandis qu'un gris clair signifiera que la scène sonore sera satisfaisante selon ce critère de qualité.

Choix des métadonnées

Ces critères doivent pouvoir évoluer en temps réel et permettre la prédiction du critère pour des échantillons qui seront joués quelques secondes après que le critère ait été représenté. Les calculs du critère ne peuvent pas être réalisés à partir de tous les échantillons de tous les signaux (à moins que ne nombre de signaux soit faible), la quantité de calculs serait alors en effet trop élevée [1]. Il est donc nécessaire de définir les metadonnées qui décriront au mieux les signaux et permettront le calcul des indices ou critères présentés (compromis entre flux à transmettre et généricité, permettant le calcul des critère). Ces metadonnées dépendent des critères que l'on souhaite représenter mais d'une façon générale, l'évolution du niveau moyen de chaque signal dans plusieurs bandes de fréquences semble suffire à la majorité des critères comme les critères de masquage. Il pourra cependant être nécessaire de rajouter des metadonnées en fonction des critères que nous souhaiterons rajouter par la suite comme des indices de corrélation mutuelle par exemple.

A l'heure du début du stage, plusieurs critères de qualité étaient déjà intégrés à cette interface de visualisation (voir [8]) : équilibre spatial, modèle de masquage basique, directivité des sources, ... (voir fig. 3.3 et 3.4)

Mapping critère / niveau de gris

L'interface de ListenSpace utilise des niveaux de gris pour représenter un certain critère de qualité, le critère étant au choix de l'utilisateur. Le problème se pose alors du mapping des niveaux de gris au facteur perceptif choisi. Il faut en effet qu'une certaine augmentation dans le critère corresponde à une impression équivalente d'éclaircissement du gris.

De nombreuses études psychophysiques ont été réalisées sur la sensibilité à différents paramètres sensoriels. Ernst Weber a été un des premiers scientifiques à s'intéresser au sujet, avec Gustav Fechner, dès 1860. L'idée générale de cette loi est que la sensation est proportionnelle au nombre de seuils différentiels perçus (JND ou Just Noticeable Difference). Ces échelles ne sont

[1] bien qu'elle dépende du critère choisi

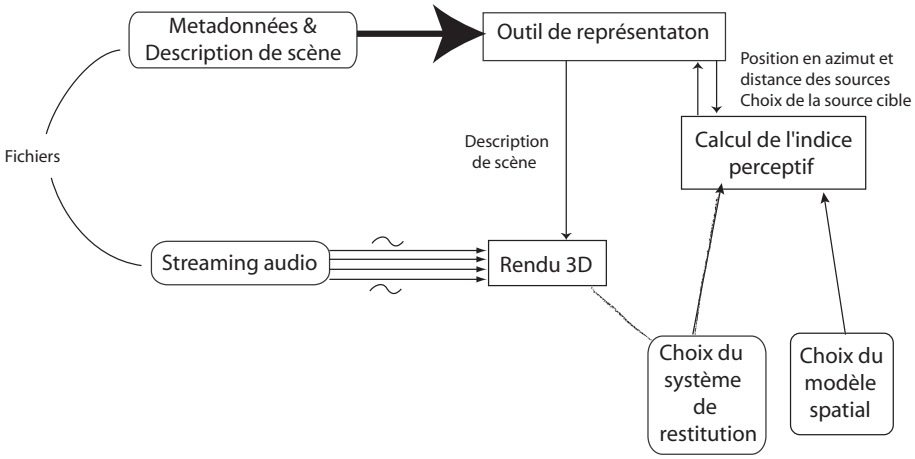

FIG. 3.3 – Schéma de fonctionnement général du système de visualisation de critères de qualité. Des fichiers audio sont lus en streaming par le moteur de rendu 3D audio. Le fichier de metadonnées est quand à lui lu par l'outil de représentation de scène sonore. Un échange entre cet outil et un algorithme de calcul permet la visualisation des critères de qualité.

pas linéaires. La correspondance entre niveaux de gris et chaque critère de qualité n'est donc pas triviale et il est nécessaire de créer une loi ou table de conversion d'un phénomène perçu à un autre. Nous ne nous sommes cependant pas consacré prioritairement à ce problème, qui reste donc à étudier de manière plus approfondie dans le futur.

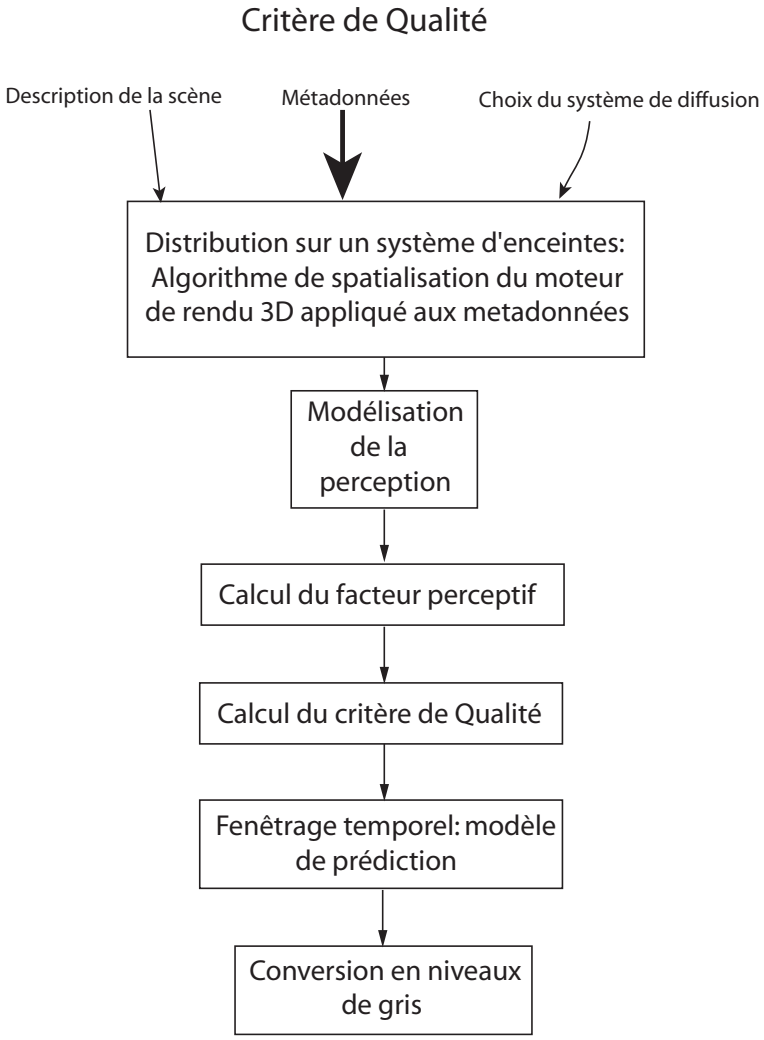

FIG. 3.4 – Schéma de calcul du critère de qualité.

Chapitre 4

Facteurs perceptifs

Le chapitre précédent traitait de l'importance de guider des utilisateurs inexpérimentés lors de spatialisation de sources sonores. Dans la partie suivante, nous nous intéressons à deux exemples de facteurs perceptifs candidats pour guider l'utilisateur : le masquage spatial et la largeur de sources fantômes. Le premier vise à représenter le risque d'une source à être masquée lorsqu'elle se rapproche en azimut ou s'éloigne en distance par rapport aux autres sources, cela à l'aide du TMR (Target-to-Masker Ratio). Le second étudie des indices objectifs permettant de prédire la largeur d'une source fantôme perçue. Nous présentons des modèles existants ainsi que des améliorations à leur apporter.

4.1 Masquage

4.1.1 Qu'est-ce que le masquage ?

Le masquage est un phénomène qui empêche un individu de percevoir un objet sonore, visuel ou autre parce qu'un autre objet a une plus grande importance perceptive. En acoustique, cela peut se manifester de différentes manières : masquage énergétique, dû à la physiologie de l'oreille , masquage temporel, lorsque deux sons sont entendus avec un faible intervalle de temps entre les deux et ce qui nous intéresse ici, le masquage spatial. En audio, le masquage est défini comme un processus via lequel le seuil d'audibilité d'un évènement sonore est élevé par un autre évènement sonore (le masqueur) [2]. Lorsque les deux évènements n'ont pas lieu en même temps, on parle alors de post-masking (masquage des sons postérieurs à un son masqueur) ou de pre-masking (masquage des sons antérieurs à un son masqueur).

L'oreille humaine est décomposée en trois parties : l'oreille externe, composée du pavillon, du conduit externe qui se finit par le tympan, l'oreille

FIG. 4.1 – Coupe de la cochlée

moyenne, composée de la trompe d'Eustache et des osselets, et enfin l'oreille interne composée de divers organes, dont un réservé à la perception auditive : la cochlée (voir fig. 4.1).

A l'intérieur de cette cochlée se trouvent de nombreuses cellules, les cellules ciliaires, qui adjointes à la membrane basilaire forment l'organe de Corti. Ces cellules agissent comme des filtres passe-bande de largeur environ un tiers d'octave ayant une réponse asymétrique. La réponse en fréquence de ces filtres varie légèrement avec la fréquence. La conséquence de ceci est qu'un son sera masqué énergétiquement si un second son a plus d'énergie dans les même bandes fréquentielles que le premier (voir fig. 4.2). Lorsque les sons ne sont pas de simples sinusoïdes mais des sons plus complexes, le problème change, dans la mesure où le cerveau humain est capable de reconstituer des sons dont il n'entend pas certaines fréquences (Psychologie de la forme).

De même, les sons ont une certaine persistance dans le cerveau : plus un son sera joué fort, plus il sera entendu longtemps après ET avant d'être joué. Un son de faible niveau joué avant ou après un son de fort niveau pourra ainsi être masqué.

4.1.2 Masquage spatial

Modèles actuels

Aujourd'hui, les études concernant le masquage spatial se concentrent principalement sur la perception des sources dans l'espace et à l'apport de cette perception par rapport à une perception monodimensionnelle. On parle d'ailleurs plus généralement de démasquage spatial que de masquage spatial, le démasquage spatial étant alors l'apport de la séparation spatiale des sources sur la perception d'une source cible. Le masquage spatial est une application des modèles de masquage fréquentiel et temporel à l'espace. Créer un modèle qui évaluerait la quantité de masquage en fonction des sources et de leur position dans la scène sonore réclame de créer un modèle de masquage perceptif du type de ceux que l'on peut trouver dans des codeurs MP3,

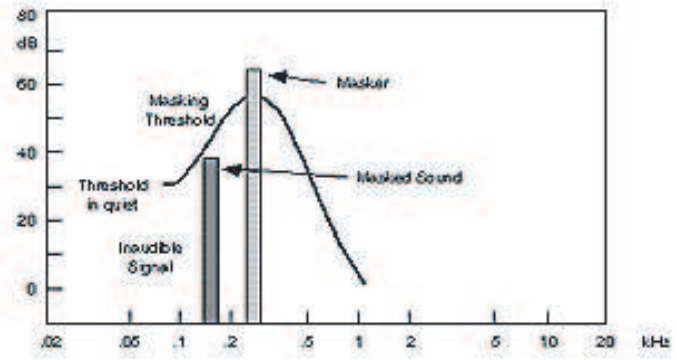

FIG. 4.2 – Deux sinusoïdes sont émises. La sinusoïde en gris clair excite des cellules ciliées dont les fréquences d'excitation sont proches de la fréquence de la sinusoïde, empêchant alors d'entendre la sinusoïde représentée en gris foncé, d'amplitude plus faible

AAC, AC3 ou WMA et d'appliquer la théorie de démasquage spatial sur ces sources. Mais en plus de l'application du masquage énergétique et temporel à l'espace, des études récentes se sont intéressées au masquage informationnel : le cerveau arrive assez bien à soustraire les bruits gênants quand ils ne contiennent pas d'information (exemple : bruit blanc). Mais dès lors que les bruits gênants contiennent de l'information, le cerveau semble avoir plus de difficultés ([14], [20], [32], [33]).

Un certain nombre de modèles de perception dans l'espace sont apparus depuis les années 50. Ils sont décrits dans [36] et respectent pour la plupart le schéma de modèle perceptif de Blauert (voir fig 4.3). Ils ont sans cesse évolué, ces dernières années ayant vu une simplification des modèles (voir [26]).

Parmi ceux-ci, citons par exemple :

- Le Modèle de l'activité du nerf auditif (ou modèle vectoriel). La première version a été mise au point par Colburn en 1973 [6]. La dernière version en date au moment de l'écriture de [11] mais un prolongement de celle-ci est en attente d'impression et a été mise au point par Ville Pulkki [15].
- Le Modèle d'égalisation et d'annulation [26].
- Le Modèle d'accumulation de Schenkel [30].
- Le Modèle de Corrélation d'Osman [25].
- Le Modèle de P.M. Zurek, décrit dans [38], et qui est le modèle dont nous nous sommes le plus servi, mais réservé à une prédiction de masquage uniquement.

Mis à part le modèle de Zurek, les autres sont des modèles complexes.

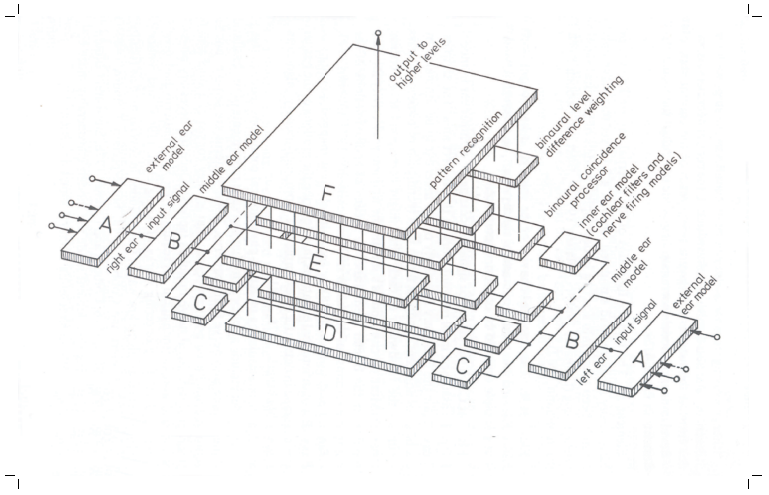

FIG. 4.3 – Diagramme fonctionnel du traitement du signal dans un modèle de localisation et de détection, [2]

Leur similitude avec le fonctionnement du cerveau en fait des modèles à priori efficaces pour étudier la perception de l'espace, mais trop gourmants en calcul.

le modèle de P.M. Zurek

Le modèle sur lequel nous avons basé la plupart de nos études est celui de P.M. Zurek. Il décompose le phénomène de masquage spatial en une partie acoustique et une partie neuronale.

Il part de la configuration décrite sur la figure 4.4 : pour un auditeur placé dans un espace anéchoïque, on teste l'intelligibilité d'une source cible S en présence d'un masqueur N. L'auditeur n'a pas le droit de bouger sa tête et on considère que la direction à laquelle l'auditeur fait face vaut 0°(voir fig. 2.2). Le niveau en dB de signal cible aux oreilles gauche et droite dans la i^{e} bande tiers d'octave est noté respectivement $S_{Li}(\theta_S)$ et $S_{Ri}(\theta_S)$. De la même manière, le niveau en dB de signal masqueur aux oreilles gauche et droite est noté respectivement $N_{Li}(\theta_N)$ et $N_{Ri}(\theta_N)$. Ces niveaux sont calculés à partir des spectres en champ libre des signaux, $S_{libre,i}$ et $N_{libre,i}$ et des HRTFs. Celles-ci sont décomposées selon les même bandes tiers d'octave, d'où on tire ensuite des niveaux $D_i(\theta)$. Les sources cible et masquante étant à la même distance, on ne tient pas compte de l'atténuation en distance. On considère les HRTFs symétriques par rapport au plan médian. Ainsi,

$$S_{Li}(\theta_S) = S_{libre,i} + D_i(\theta_S) \quad (4.1)$$

$$S_{Ri}(\theta_S) = S_{libre,i} + D_i(-\theta_S) \quad (4.2)$$

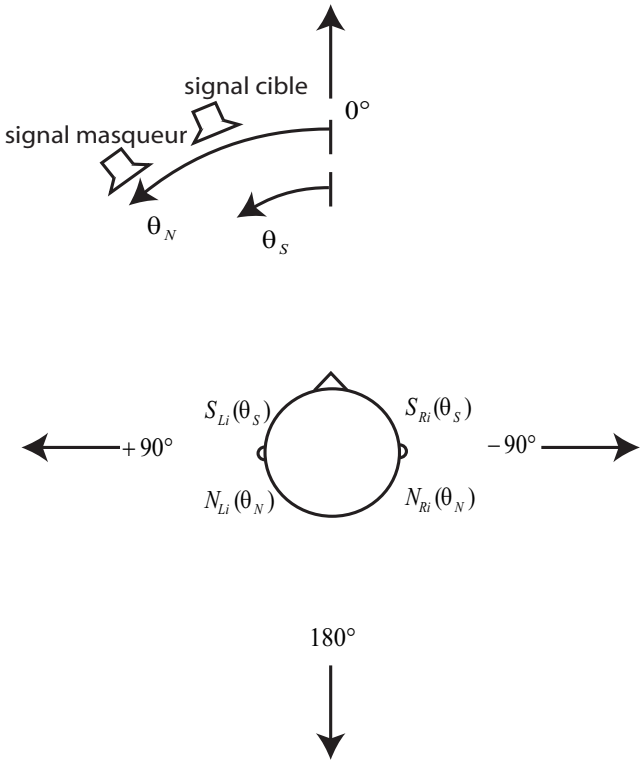

FIG. 4.4 – Le sujet écoute un signal cible et un signal masqueur, chacun étant placé à un azimut différent (θ_S et θ_N). On en déduit des niveaux aux oreilles $S_{Li}(\theta_S)$, $S_{Ri}(\theta_S)$, $N_{Li}(\theta_N)$ et $N_{Ri}(\theta_N)$

$$N_{Li}(\theta_N) = N_{libre,i} + D_i(\theta_N) \quad (4.3)$$
$$N_{Ri}(\theta_N) = N_{libre,i} + D_i(-\theta_N) \quad (4.4)$$

A partir de ces niveaux, on calcule le *Target-to-Masker Ratio*, ou TMR. On considère en effet que l'indice le plus important pour l'estimation de la quantité de masquage dans une configuration donnée est le rapport signal cible sur signal masqueur. On le calcule pour les deux oreilles et on conserve celui de la "meilleure oreille", i.e. celle qui a le meilleur TMR. Le cerveau semble en effet pouvoir dans une certaine mesure se concentrer sur cette meilleure oreille pour extraire l'information qu'il désire (ce qui explique partiellement l'effet *cocktail party*, dont traite [14]). A partir de cet indice, on peut déduire un niveau d'intelligibilité, comme nous allons le voir plus tard.

Les équations précédentes nous permettent de déduire le TMR à l'oreille gauche :

$$TMR_{L,i} = S_{Li}(\theta_S) - N_{Li}(\theta_N)$$
$$TMR_{L,i} = S_{libre,i} - N_{libre,i} + D_i(\theta_S) - D_i(\theta_N) \quad (4.5)$$

De même, on obtient le TMR à l'oreille droite :

$$TMR_{R,i} = S_{Ri}(\theta_S) - N_{Ri}(\theta_N)$$

$$TMR_{R,i} = S_{libre,i} - N_{libre,i} + D_i(-\theta_S) - D_i(-\theta_N) \tag{4.6}$$

(Les bandes de fréquence utilisées sont les 15 bandes centrées en $f_1 = 200, \ldots, f_{15} = 5000$ *Hz, selon les tables de Shawn et Vaillancourt).*

On peut alors appliquer un modèle de prédiction d'indice d'intelligibilité standard (ANSI, 1969, méthode tiers d'octave [16]). On peut simplifier le modèle si on fait quelques hypothèses : le niveau de bruit est bien en dessous du seuil d'audition et le spectre du signal masqueur n'a pas de pente raide, de manière à ce qu'aucun effet de out-of-band n'apparaisse : si un des signaux a un contenu fréquentiel avec une pente raide, par la méthode de calcul de l'indice d'articulation, il aura tout de même de l'énergie dans des bandes adjacentes (parce qu'on utilise des filtres ERB -Equivalent Rectangular Bandwidth-, plus proche du fonctionnement de nos oreilles). Lorsque ces conditions sont remplies, alors l'indice d'articulation ne dépend plus que des rapports TMR dans les différentes bandes fréquentielles. Pour l'oreille gauche, l'indice d'articulation IA vaut

$$IA_L = \sum_{i=1}^{15} w_i \gamma_{Li} \tag{4.7}$$

avec

$$\gamma_{Li} = \begin{cases} 0 & \gamma_{Li} < -12 \\ \gamma_{Li} + 12 & -12 \leq \gamma_{Li} \leq 18 \\ 30 & \gamma_{Li} > 18 \end{cases}$$

γ_{Li} est alors l'"efficacité de la bande" et w_i est le "poid d'importance fréquentielle", comme indiqué dans la norme ANSI S3.5-1969. L'indice d'articulation sur l'oreille droite est calculé de la même manière. Des résultats de pourcentage de reconnaissance en fonction de l'indice d'articulation sont donnés dans cette même norme.

Il s'agit là de la prédiction d'écoute monorale. Nous nous intéressons dans ce projet à une écoute binaurale [1] et à l'apport de ce mode d'écoute sur le monoral. Zurek pose deux hypothèses sur l'écoute binaurale.

La première consiste à dire que dans une bande tiers d'octave, en écoute binaurale, le TMR effectif n'est jamais pire que celui de la meilleure oreille monorale. Pour chaque bande, on prend donc le TMR de la meilleure oreille. Etant donné le comportement global des HRTFs, on peut supposer que presque tous les TMRs, dans une configuration donnée, seront pris sur la même oreille.

[1] dans le sens où on écoute avec deux oreilles, et non dans le sens du système de prise de son / synthèse / restitution décrit dans la partie 2.1.1

La seconde hypothèse est que le TMR effectif peut être amélioré (mais pas détérioré) par une écoute binaurale. En pratique, pour la i^{e}bande, on calcule le MLD (Masking Level Difference ; il s'agit ici de l'apport d'un traitement psychophysique des signaux des deux oreilles tel qu'il a été décrit par Colburn dans [6]) pour un jeu de paramètres interauraux. Ce nombre de décibels (toujours positif) est ajouté au meilleur TMR mesuré précédemment.

$$TMR_{B,i} = max(TMR_{R,i}, TMR_{L,i}) + MLD_i \tag{4.8}$$

Le calcul du MLD_i se fait d'après la théorie de Colburn [6] sur la perception dans l'espace selon un modèle de nerf auditif.

$$MLD_i = 5log_{10}\{1 + \frac{C_i}{M_i}[1 + \beta_i^2 - 2\beta_i cos(\Phi_{Si} - \Phi_{Ni})]^2\} \tag{4.9}$$

où

$$\begin{aligned}
C_i &= \text{constante qui détermine la dépendance en fréquence du MLD} \\
\beta_i &= \alpha_{Si}/\alpha_{Ni} \\
\alpha_{Si} &= 10^{(S_{Li}-S_{Ri})/20}\text{, rapport interaural de niveaux RMS} \\
&\quad \text{de signal cible} \\
\alpha_{Ni} &= 10^{(N_{Li}-N_{Ri})/20}\text{, rapport interaural de niveaux RMS} \\
&\quad \text{de signal masqueur} \\
M_i &= max(1, \beta_i^4) \\
\Phi_{Si} &= \begin{cases} 2\pi f_i \tau_i(\theta_S) & |\theta_S| \leq \pi/2 \\ 2\pi f_i \tau_i(\pi - |\theta_S|)sin(\theta_S) & |\theta_S| \geq \pi/2 \end{cases} \\
\Phi_{Ni} &= \text{même chose, mais en remplaçant } \theta_S \text{ par } \theta_N \\
f_i &= \text{fréquence centrale de la i}^{\text{e}}\text{ bande tiers d'octave} \\
\tau_i(\theta) &= \begin{cases} \frac{3r_l}{c} sin\theta & i = 1-5 \\ (3 - \frac{log_2(f_i/500)}{2})\frac{r_l}{c} sin\theta & i = 6-9 \\ \frac{r_h}{c}(\theta + sin\theta) & i = 10-15 \end{cases}
\end{aligned}$$

avec

$$\begin{aligned}
r_l &= \text{9.3cm, rayon de la tête nominal en basses fréquences} \\
r_h &= \text{8.75cm, rayon de la tête nominal en hautes fréquences} \\
c &= \text{344m/sec, vitesse du son dans l'air}
\end{aligned}$$

Les C_i peuvent êtres calculés à partir de mesures sur le schéma 15.2 de [38]. On déduit alors l'indice d'articulation de la même façon que dans un cas monoral à partir du $TMR_{B,i}$ et on obtient

$$IA_B = \sum_{i=1}^{15} w_i \gamma_{Bi} \tag{4.10}$$

avec

$$\gamma_{Bi} = \begin{cases} 0 & \gamma_{Li} < -12 \\ \gamma_{Bi} + 12 & -12 \leq \gamma_{Bi} \leq 18 \\ 30 & \gamma_{Bi} > 18 \end{cases}$$

Au delà de ce travail, il semblerait que de nombreuses études qualitatives aient été effectuées ([14], [31], [33], [34]), s'appuyant toutes sur un des points du modèle de Zurek qu'il faudrait améliorer. Les modèles actuels ne tiennent pas compte du champ diffus. Quelques publications ont cependant été faites à ce sujet [24].

les cartes de saillance

En audio comme en traitement de l'image, on s'intéresse depuis longtemps au masquage. En revanche des approches différentes ont été abordées dans les deux domaines. En traitement de l'image, on s'intéresse à ce que l'on appelle les cartes de saillance : savoir ce qui va ressortir dans une image selon différents critères : couleur, intensité lumineuse, orientation des objets graphiques, différences centre-contours, ...

Pour cela, on passe notre image à travers différents filtres (par exemple passe-haut vertical, passe-haut horizontal ou bien passe-bas) à différentes résolutions, puis on recombine les différentes sous-images obtenues en une carte de saillance (saliency map) qui met en valeur les objets qui attireront le plus notre œil. Très récemment, des expériences ont tenté de rapprocher cette méthode de la perception audio, afin de savoir ce qui ressort d'une scène sonore. Les résultats sont semble-t-il intéressants et se rapprochent plus de la réalité qu'une simple analyse d'intensité [22].

Le prototype d'Auditory Saliency Map proposé par Kayser et al. part d'un son audio monophonique composé d'un son cible et d'un bruit blanc. Le but de ce projet est de savoir si ce son cible ressort ou non de la scène. Les auteurs commencent par effectuer une TFCT à échelle linéaire sur le son monophonique. Il extraient ensuite différents paramètres du spectrogramme obtenu à l'aide de la TFCT : intensité, contraste fréquentiel (dérivée verticale du spectrogramme) et contraste temporel (dérivée horizontale du spectrogramme), et ce à différentes échelles. Ils recombinent ces images de paramètres obtenus pour en déduire cette carte de saillence auditive.

Cette carte est censée indiquer ce qui ressortira du signal monophonique. Cependant, quelques réserves sont à porter sur ce projet : l'échelle fréquentielle utilisée dans ce projet est linéaire, aucun coefficient de pondération perceptive sur le plan fréquentiel est appliqué et bien que les auteurs semblent satisfaits de leur modèle, la corrélation entre les résultats du modèle et ceux donnés par les sujets semble assez faible (0.47 en moyenne). La corrélations entre ces résultats est en revanche toujours améliorée par le passage à un

modèle d'intensité pure à un modèle utilisant plusieurs indices (comme les contrastes fréquentiels et temporels).

Nous avons donc décidé de nous pencher sur cette idée de saillance afin d'améliorer le modèle actuel de Zurek tout en corrigeant certains points : une échelle fréquentielle adaptée et un modèle d'intensité plus performant et plus complet, celui de P.M. Zurek.

4.1.3 Améliorations à apporter à ces modèles

Comme l'ont très bien montré les récentes publication de Kidd et al. et de Shinn-Cunningam et al., de nombreux points sont à améliorer sur le modèle qui actuellement est considéré comme le plus performant dans l'évaluation de la quantité de masquage, le modèle de Zurek [38].

Masquage spatial par un champ diffus

Les expériences sur le masquage de la parole ont souvent été réalisées en champ libre. Dans des conditions courantes, le champ diffus est bel et bien présent. Il contribue lui aussi à masquer la source cible... qu'il s'agisse du champ diffus d'une source masquante ou bien de celui de la source cible. Ce problème est abordé dans [12], [31] et [39].

Une solution simplifiée pourrait être, comme dans [23], de rajouter un handicap dû à la réverbération, sous forme d'un nombre de dB à soustraire au $\gamma_{B,i}$ de Zurek. Il faudrait pour cela réaliser des tests sur l'handicap que procure un champ diffus selon que la source en champ diffus soit la source cible ou bien une ou plusieurs sources masquantes, ainsi que suivant le type de réverbération et les directivités des sources.

Système de restitution

Le modèle de Zurek n'est fait que pour une écoute directe des sources, sans passer par un système de restitution. Il est vraisemblable qu'une source ne sera pas masquée de la même façon suivant que la restitution se fasse en binaural ou bien en mono. Il est donc nécessaire de tenir compte du système de diffusion (voir la partie 2.1.1). Cela fait partie des améliorations que nous avons apportées au modèle au cours de ce stage.

Modèle de masquage en champ proche

Il faut ou non améliorer le modèle suivant que le système de restitution utilise ou non un algorithme de masquage en champ proche réaliste. Le modèle de masquage de Zurek ne tient pas compte de la distance, mais le modèle de masquage initial intégré à l'outil de représentation de facteurs perceptifs de ListenSpace l'intègre aux équations de manière simplifiée : le niveau global de la source décroit en $\frac{1}{d}$. Ce modèle convient tout à fait à des sources

placées à plus d'un mètre de l'utilisateur, mais pas pour des sources plus proches : ce modèle considère que mis à part l'atténuation dûe aux HRTFs, le niveau de signal dans les deux oreilles est le même. Or pour des faibles valeurs de d, la distance, et donc l'atténuation dûe à cette distance, n'est plus du tout la même d'une oreille à l'autre, excepté dans le plan sagital. Nous avons choisi de prendre la même distance pour les deux oreilles dans notre modèle. Cependant, Blauert ayant donné dans [2] des informations nous permettant de calculer cette différence de canal acoustique, nous avons créé une fonction qui calcule si on le souhaite la distance de la source à chacune des deux oreilles en fonction de la distance d au centre de la tête et de l'azimut az :

Dans le cas où $d > 1$

$$dist_{min} = d - \frac{D}{2}\ \sin(|az|) \tag{4.11}$$

$$dist_{max} = d + \frac{D}{2}\ az \tag{4.12}$$

où D est la distance entre les deux oreilles, soit en général 17cm.
Dans le cas où $d \leq 1$

$$dist_{min} = D\ \sqrt{n^2 + n + 0,5 - (n + 0,5)\sin(|az|)} \tag{4.13}$$

$$dist_{max} = D\ ((n + 0,5)\ \cos(e) + 0,5(az + e)); \tag{4.14}$$

avec $n = \frac{d - D/2}{D}$ et $e = \arcsin(\frac{D}{2\ d})$
D'après [34], certaines configurations spatiales peuvent amener une grande quantité de démasquage spatiale dont on ne s'apercevrai pas si on ne modélisait pas cette différence d'intensité au niveau des oreilles en champ proche.

Modèle de masquage informationnel

Le modèle de Zurek est un modèle de masquage énergétique avec traitement psychophysique de l'information (perception dans le sens bottom-up, soit sans retour du cerveau). Le masquage informationnel est en cours d'étude ([9], [12], [13], [14], [20], [29], [33]) mais pour le moment, encore aucun modèle n'existe...

Masquage temporel lorsque deux sources sont cohérentes

Si deux sources n'ont qu'un décalage temporel global, une différence d'intensité globale ou une combinaison des deux, on considère ces deux sources comme cohérentes. Le masquage d'une source par une autre dans ce cas est fortement lié au retard et à la différence de niveau entre les deux signaux. On peut constater trois différents cas de perception, lorsqu'il y a deux sources

cohérentes : soit on ne perçoit qu'une seule source à l'endroit d'une des deux sources réelles, auquel cas la seconde est masquée, soit on perçoit une source entre les deux sources réelles, auquel cas les deux sources contribuent à la perception, soit on perçoit deux sources distinctes aux lieus des deux sources réelles. Les phénomènes de masquage et de formation d'image fantôme sont donc fortement liés. C'est pourquoi nous avons été amené à étudier ce dernier également.

4.2 Images fantômes

De nombreux projets ont étudié l'apparition de sources sonores virtuelles en fonction des sources réelles émettant des ondes acoustiques. Une partie de ces études est résumée par Jens Blauert dans [2] : si les procédés qui rentrent en jeu lors de la perception dans l'espace d'une source seul sont principalement physiques, l'apparition d'une seconde source sonore peut engendrer différentes phénomènes perceptifs (voire même des illusions) qui dépendent principalement de la fréquence des sons et de leur similarité en temps et en intensité.

Pour étudier la similarité entre deux signaux, on calcule leur corrélation normalisée.

$$\phi_{s_1\,s_2}(\tau) = \lim_{T\to\infty} \frac{\frac{1}{2\,T}\int_{-T}^{+T} s_1(t)s_2(t\ +\ \tau)\ \mathrm{d}t}{\frac{1}{2\,T}\sqrt{\int_{-T}^{+T} s_1^2(t)\ \mathrm{d}t \int_{-T}^{+T} s_2^2(t)\ \mathrm{d}t}} = \frac{\overline{s_1(t)s_2(t\ +\ \tau)}}{s_{1_{rms}} s_{2_{rms}}} \quad (4.15)$$

4.2.1 Deux sources sonores cohérentes

On qualifie de cohérentes deux sources pour lesquelles

$$\max_{\tau}\ |\phi_{s_1\,s_2}(\tau)|\ =\ 1. \quad (4.16)$$

On considère donc ici que deux signaux sont cohérents si ils sont identiques à un retard global $\delta\tau$ ou à un gain global δL près.

On peut dans ce cas distinguer trois phénomènes différents :

- L'auditeur entend une seule source sonore, virtuelle, placée entre les deux sources sonores réelles. On dit alors que les sources s1 et s2 sont globalement cohérentes et qu'aucune ne masque l'autre.
- L'auditeur entend une seule source sonore, mais localisée à la place d'une des deux sources s1 et s2. On dit alors que l'une des deux sources masque l'autre.
- L'auditeur entend deux sources localisées aux lieux des sources s1 et s2. On dit alors que les deux sources sont incohérentes.

La majeure partie des tests de perception concernant la formation d'images fantômes à partir de deux sources cohérentes ou non ont été réalisées à l'aide

de casques audio. Le cerveau situe la source sonore en fonction de la différence de temps entre les deux sources et de leur différence d'amplitude, la tête faisant office en général d'obstacle acoustique et atténuant donc le niveau d'un son d'une oreille à l'autre. [2]

En revanche, des travaux comme ceux de Mike Williams [37] nous permettent de déduire des règles de perception des sources en fonction de l'espacement entre les microphones, de l'orientation des capsules et de la position de la source.

Mais le coefficient d'intercorrélation $k = \max_\tau \ |\phi_{s_1\, s_2}(\tau)|$ ne vaut pas toujours 1. Si $k < 1$, on parle alors de sources partiellement cohérentes.

4.2.2 Deux sources partiellement cohérentes

Une règle générale doit être énoncée ici : le degré de cohérence de signaux est différent au niveau des sources et au niveau des oreilles d'un sujet. Deux sources cohérentes entraînent généralement des signaux aux oreilles fortement cohérents, mais pas complètement (l'influence de l'acoustique du lieu dans lequel on se trouve, principalement). De même, si deux sources sont incohérentes dans la nature, les signaux aux oreilles seront tout de même partiellement cohérents[2].

Des tests ont été effectués par Chernyak et Dubrovky [5] et mettent en avant l'influence de la cohérence des signaux aux oreilles sur la localisation de sources fantômes. Il a alors été constaté que pour un coefficient d'intercorrélation valant 1, le sujet perçoit une source centrée et quasi-ponctuelle, localisée entre les deux sources réelles. Plus le coefficient décroit, plus la source perçue est large. Pour un coefficient vararіant entre 0,4 et 1, la source reste unique. En dessous de 0,4, la source se scinde et le sujet perçoit alors deux sources distinctes sur ses oreilles.

Le degré de cohérence a également une influence sur le flou de localisation : une image fantôme sera moins précisément localisée si les deux sources qui créent cette image ne sont pas cohérentes [17].

A la suite des travaux que nous avons réalisé sur le masquage, nous nous sommes également penché sur le problème de perception d'artefacts lors de la spatialisation de sources démixées.

4.3 Perception d'artefacts lors de spatialisation de sources démixées

Une fois un mixage réalisé et stocké sur un support dans un format donné (stéréo, binaural, 5.1, ...), il est souvent difficile de retoucher ce mixage aussi

[2] c'est d'ailleur cette problématique que le système de diffusion Transaural tente de résoudre

librement qu'on le souhaiterait, puisqu'on ne dispose plus des sources originales. Pour pallier ce problème, il a été mis au point de nombreux algorithmes de séparation de sources (analyse en composantes indépendantes, filtrage de Wiener, modèlisation des sons, ...), permettant ainsi d'extraire d'un mixage les sources originales de façon plus ou moins fructueuse... Il peut ainsi apparaître différents types de bruits de fond, interférences, artefacts,...

L'intérêt grandissant pour ces algorithmes soulève la question de qualité des sources extraites. Si des méthodes d'évaluation de tels algorithmes ont été mises au point [10], elles ne tiennent pas compte de la spatialisation des sources. Il a en effet été constaté que lorsque l'on écoute une source monophonique constituée de la superposition de deux sources démixées, on perçoit peu ou pas du tout d'artefacts de séparation de source. Le bruit est en effet négligeable et les artefacts et autres interférences de chaque signal extrait se recombinent efficacement avec les autres signaux. En revanche, si on déplace dans l'espace une source extraite des autres, on peut percevoir des artefacts. Nous avons souhaité nous intéresser à la prédiction de cette perception d'artefacts. Il pourrait en effet être intéressant de représenter un critère de qualité basé non plus sur le masquage ou la largeur des sources sonores, mais sur la perception d'artefacts.

L'idée de ce critère de qualité est comme précédemment de s'intéresser à ce que perçoit l'utilisateur, et pour cela, aux signaux qui arrivent à ses oreilles. Pour le moment, nous supposons que nous disposons à la fois des signaux issus de la séparation de source et des signaux originaux non mixés. On recherche un critère de qualité qui soit optimal pour les signaux originaux.

L'étude que nous avons réalisée consistait à spatialiser deux sources (soit les sources originales $s1$ et $s2$, soit celles issues d'un algorithme d'extraction de scène sonore basé sur une filtrage de Wiener [28], $s1_{estim}$ et $s2_{estim}$) en les séparant d'un angle 2α autour de l'axe d'azimut 0. L'algorithme de spatialisation est une combinaison linéaire des deux sources spatialisées.

$$L = Is1 + Cs2 \tag{4.17}$$

à l'oreille gauche et

$$L_{estim} = Is1_{estim} + Cs2_{estim} \tag{4.18}$$

à l'oreille droite.

Le gain I ou C est celui des HRIR correspondant à la position spatiale que l'on souhaite simuler. Par souci de précision, les signaux filtré par les 15 filtres ERB décrits dans la partie 4.1.2. De même, les gains des HRIR sont donnés dans chaque bande de fréquence. I est le gain correspondant au trajet ipsilatéral (d'une source à l'oreille qui lui est la plus proche) et C celui correspondant au trajet contralatéral (de la source à l'oreille qui est la plus éloignée). On suppose les HRIRs symétrisées, c'est-à-dire que la HRIR pour

l'oreille gauche correspondant à un angle α est la même que celle de l'oreille droite pour un angle $-\alpha$.

On dispose d'un outil d'évaluation d'algorithmes d'extraction de scène sonore décrit dans [4]. Cet outil évalue, pour un signal extrait donné, et les signaux originaux (avant mixage par un éventuel ingénieur du son) qui ont servi à réaliser l'exemple d'extraction, la partie s_{target} du signal extrait qui est corrélée avec le signal original cible, la partie e_{interf} de ce signal qui est corrélée avec les autres sources mais pas avec la source originale cible, et enfin la partie e_{artef} qui n'est corrélée avec aucune des sources originales : $s_{estim} = s_{target} + e_{interf} + e_{artef}$. Nous souhaitons donc appliquer cet outil à nos signaux spatialisés : on choisi pour signal démixé L_{estim} et pour signaux originaux L et R. A partir de ces signaux, on peut déduire un rapport signal/bruit

$$SDR = \frac{||s_{targetleft}||^2}{||e_{interf} + e_{artef}||^2} \tag{4.19}$$

Un cas particulier est cependant à noter : si l'angle α entre les deux sources est nul (les deux sources sont alors positionnées devant l'utilisateur), L et R sont colinéaires. e_{interf} ne peut alors pas être défini.

Nous n'avons cependant pas eu suffisamment de temps pour développer cette problèmatique autant que nous l'aurions souhaité. Si nous avons effectivement mené une réflexion sur le problème, elle nous a semblé trop juste pour paraître dans ce rapport.

Si ces réflexions amènent des pistes de travail dans le cadre de ce stage, nous avons tout de même préféré nous concentrer sur des méthodes de prédiction de niveau de masquage. Nous pouvons néanmoins imaginer un travail futur sur la perception d'artefacts lors du mixage de sources démixées, sur la localisation des sources et sur le détimbrage pour l'aide au mixage.

Chapitre 5

Implémentation

Dans la partie précédente, nous avons présenté un algorithme de masquage ainsi que les améliorations à lui apporter dans le cadre de ce projet. Nous développons ici les différentes implémentations réalisées sous Matlab et ListenSpace. Le chapitre commence par une présentation des fonctions dont on disposait initialement puis continue par la boîte à outils que nous avons développée, aussi bien pour un système de diffusion binaurale que pour un système stéréo. Cette partie ne se concentre que sur le masquage, l'implémentation d'autres algorithmes perceptifs étant réservée à des études ultérieures.

Le but final de ce stage était d'implémenter une représentation de facteurs perceptifs sous ListenSpace, utilisé en parallèle de MAX/MSP et du SPAT de l'IRCAM. Pour cela, nous avons procédé en deux étapes :

- Mise en place de l'algorithme et tests préliminaires sous Matlab
- Implémentation sous ListenSpace et MAX/MSP en minimisant la quantité de calculs à effectuer en temps réel

Les différentes étapes de calcul sont expliquées sur le schéma 5.1

5.1 Ce dont on disposait initialement

En plus de Matlab et de ses fonctions de base, nous disposions également d'une Auditory Toolbox développée par Malcom Slaney (Apple Computer et Interval Research Corporation), qui comprenait de nombreuses fonctions liées à la perception auditive. Parmi ces fonctions, celles qui nous ont été les plus utiles sont deux fonctions permettant de générer un jeu de filtres ERB (Equivalent Rectangular Bandwidth, similaires à des filtres en bandes de Bark, mais au comportement plus proche de celui de l'oreille humaine) et de filtrer un signal par ces filtres, récupérant ainsi les signaux perçus dans

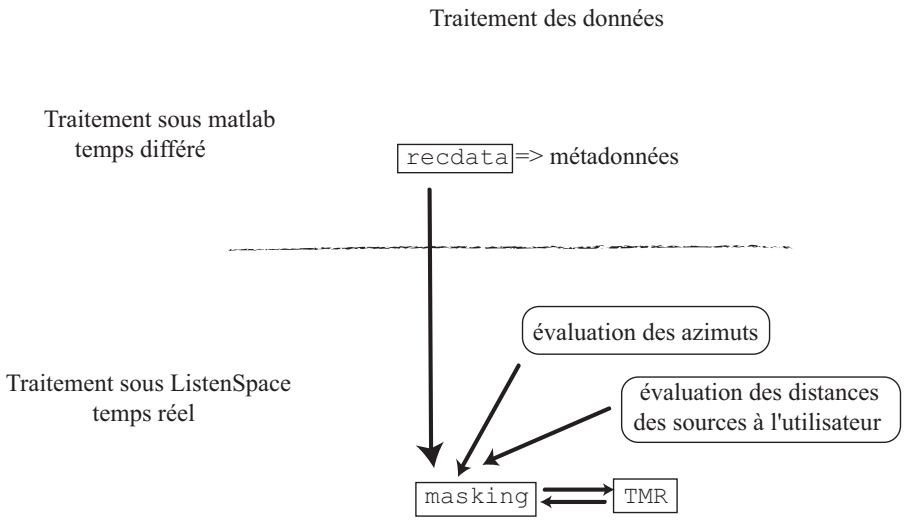

FIG. 5.1 – Organisation des calculs pour l'évaluation du critère de masquage. Les signaux doivent être traités préliminairement par la fonction `recdata` pour en extraire les métadonnées. Ensuite, suivant que l'on souhaite réaliser le traitement en temps réel (sous ListenSpace, avec un moteur de rendu et un outil d'envoi des métadonnées) ou en temps différé (sous Matlab, qui s'occupe alors lui-même du rendu sonore en plus de l'évaluation de masquage), les métadonnées sont soit envoyées à ListenSpace, soit à Matlab. Le traitement alors réalisé dans l'un ou dans l'autre est très similaire. Ce calcul des métadonnées, qui ne dépendent que des signaux sources et non de la spatialisation ou du système de restitution, permet une paramétrisation complète du critère de qualité en temps réel. Cela est rendu possible par la relative simplicité des calculs à faire en temps réel (pas de convolution, principalement des gains à appliquer).

chaque bande critique. Nous disposions également des jeux de HRTFs de la banque LISTEN, enregistrés à l'IRCAM. Ces jeux de HRTFs proposent, pour chaque individu dont les HRTFs ont été enregistrées, 187 réponses impulsionnelles enregistrées dans différentes positions autour de la tête.

Nous avons commencé par créer un jeu des fonctions de base dont nous allions avoir besoin. Elles nous permettent d'évaluer la quantité de masquage dans diverses configurations pour le cas où nous aurions au maximum trois sources , l'algorithme étant facilement extensible au-delà.

5.2 Fonctions version binaural

Une fonction `bino` prend les bases de HRTFs calculées, un signal, un azimut et les distances de la source à chacune de oreilles en argument. La fonction renvoie les deux canaux du downmix binaural de la source en sortie.

Une fonction `dist` calcule la distance de la source à chacune des oreilles. En effet, il a été montré dans [35] et [21] que la distance entre les deux oreilles pouvait jouer un rôle important dans la perception de sources en champ proche. Note : il a été décidé que le champ proche n'était pas une priorité et que nous n'avons donc finalement pas utilisé cette fonction.

Une fonction `hrir_build` qui réorganisait les structures de données de HRTFs pour que nous puissions l'utiliser plus facilement. La fonction ne conserve que les HRIR correspondant à des positions dans le plan azimutal, puisque c'est dans ce plan que nous travaillons en général, les filtre selon les filtres ERB puis stocke les moyennes de chaque bande et de chaque HRIR dans une matrice.

5.2.1 recdata

La fonction `recdata` met en mémoire les données nécessaires et réalise tous les calculs qui ne dépendent pas des positions des sources. `recdata` commence par fixer plusieurs variables : le nombre `nbandes` de bandes ERB que l'on va générer, la longueur `l_fenetre` de la fenêtre de moyennage ainsi que le nombre d'échantillons `f_update`entre le début d'une fenêtre et celui de la suivante (cf moyennage temporel ci-après).

La fonction charge les fichiers audio et les tronque (par souci de simplicité, nous avons pris des fichiers de même longueur). Il calcule ensuite les coefficients des filtres ERB puis charge les HRTFs et calcule leurs moyennes à l'aide de `hrir_build`. Nous avons choisi de prendre le jeu compensé et non brut. Un jeu de HRTFs compensées est un jeu de HRTF normalisé par la HRTF de champ diffus (moyenne des HRTFs pondérées par l'angle solide qu'elles représentent), afin de compenser la réponse du micro et du haut-parleur.

Pour chaque direction et chaque bande, on calcule le gain à appliquer aux signaux :

$$gainhrir_{left,i,az} = 20\log(\sqrt{\sum((hrir_{left,az} * ERB_i)^2)}) \tag{5.1}$$

$$gainhrir_{right,i,az} = 20\log(\sqrt{\sum((hrir_{right,az} * ERB_i)^2)}) \tag{5.2}$$

ici, i est le numéro de bande et az est l'azimut correspondant à la HRIR. Ces gains sont consignés dans une base de données qui pourra être utilisée "à la volée" pour calculer les indices de masquage dans ListenSpace.

On filtre ensuite chacun de nos signaux audio par chaque filtre ERB :

$$y_i = \text{fft}^{-1}(X\text{x} \quad \text{ERBfft}_i) \tag{5.3}$$

avec $X = \text{fft}(x)$ et $\text{ERBfft}_i = \text{fft}(ERB_i)$

On crée ensuite une fenêtre de moyennage `fen` du signal, puis on calcule tous les $N_{overlap}$ échantillons, pour chaque canal de chaque signal original, une moyenne $Y_i(N)$ avec

$$Y_i(N) = 20\log\left(\sqrt{\sum_{n=0}^{L_{fen}-1}(\text{fen}(n)y_i((N-1)N_{overlap}+n+1))^2}\right) \tag{5.4}$$

Note : dans ce procédé, on perd de l'information et on ne peut pas reconstituer exactement le signal original. La même méthode que celle appliquée dans l'algorithme décrit ici a été utilisée sur des bancs de filtres avec succès. Le principal inconvénient de cette méthode était la réaction face aux différentes fréquences : il traitait de manière similaire toutes les fréquences et la pente des filtres passe-bande n'était pas adaptée, faussant grandement les résultats. En effet, dans le banc de filtres que nous utilisions, tous les filtres avaient la même largeur de bande. Il a donc été décidé d'utiliser des filtres ERB, ne permettant certes pas de reconstituer le signal original de manière exacte, mais se rapprochant beaucoup plus de notre perception et donnant de bien meilleurs résultats.

Le programme se débarrasse ensuite des signaux des bandes de fréquence centrale supérieure à 6 kHz. En effet, la bande de fréquence la plus haute traitée par Zurek est centrée autour de 5 kHz, puisqu'il étudie la voix. Nous nous intéressons à des signaux musicaux, qui peuvent monter bien au-delà de 5 kHz. Mais les tests que nous avons effectué ont montré qu'il était suffisant de se limiter à 5 kHz pour les problèmes de masquage. Ceci pourrait venir des signaux utilisés (voir fig. 6.1).

5.2.2 TMR

Cette fonction prend en arguments d'entrée les $Y_{i,left}$ et $Y_{i,right}$ correspondant à chaque source initiale. Elle calcule un $Y_{i,left,masker}$, un $Y_{i,left,target}$

et leurs équivalents en signaux à l'oreille droite.

$$\begin{cases} Y_{i,left,target} = Y_{i,left} \text{ de la source cible} \\ Y_{i,left,masker} = 20\log\left(\sum_{Y_{i,left}} \text{des autres sources } 10^{Y_{i,left}/20}\right) \end{cases} \quad (5.5)$$

Les calculs sont similaires pour l'oreille droite.

On déduit ensuite les rapport signal/bruit à chacune des oreilles :

$$SNLeft_i = Y_{i,left,target} - Y_{i,left,masker} \quad (5.6)$$

$$SNRight_i = Y_{i,right,target} - Y_{i,right,masker} \quad (5.7)$$

Le programme calcule ensuite pour chaque bande le meilleur Target-to-Masker Ratio :

$$TMR_i = max(SNLeft_i, SNRight_i) \quad (5.8)$$

5.2.3 masking

C'est notre fonction principale. C'est elle que l'utilisateur lance quand il veut tester une configuration de scène sonore en particulier et savoir si il pourra entendre une source donnée et si oui, avec quel degré de qualité.

Cette fonction prend en arguments :

- Un vecteur `d` de distances, spécifiant les distances auxquelles se trouvent chacune des trois source
- Un vecteur `az` spécifiant les azimuts des sources (en degrés) entre -180 et 180 degrés
- Pour chaque source, à la fois le vecteur de signal original x et la matrice de moyennes temps-fréquence Y_i telle qu'elle a été calculée par `recdata`.

Elle renvoie en sortie

- Le signal binauralisé correspondant à la scène sonore décrite
- Le TMR final (i.e. celui calculé par la fonction `TMR` puis modifié par le Masking Level Difference)
- Le Masking Level Difference (MLD)
- les TMR gauche et droite calculés par la fonction `TMR`

La fonction commence par sauvegarder les azimuts dans une nouvelle variable locale `azold` afin de pouvoir les réutiliser pour le calcul du Binaural Masking Level Difference. On calcule l'indice des HRTFs que l'on va utiliser pour chaque source (puisque les HRTFs ont été mesurés dans le plan azimutal tous les 15 degrés, ind_az = az/15 + 1).

La fonction calcule ensuite la portion de chaque signal qui arrive à chacune des oreilles de l'auditeur. Pour un signal donné, on obtient ainsi

$$Y_{i,left} = Y_i - 20\log(d_{source}) + gainhrir_{left,i,az_{source}} \quad (5.9)$$

$$Y_{i,right} = Y_i - 20\log(d_{source}) + gainhrir_{right,i,az_{source}} \quad (5.10)$$

Il s'agit là d'un modèle de spatialisation simple. D'autres modèles peuvent être utilisés pour tenir compte, par exemple, de l'effet de salle.

où d_{source} et az_{source} sont respectivement la distance de la source à l'utilisateur et l'azimut de la source. On atténue donc chaque source en fonction de sa distance (en considérant la source comme sphérique) et de son azimut (en tenant compte des HRTFs choisis).

Une fois tous ces signaux calculés, il est possible d'utiliser la fonction TMR décrite précédemment afin de calculer un Target-to-Masker Ratio à la meilleure oreille.

Il ne reste "plus qu'à" calculer le MLD. Pour cela, on souhaite appliquer simplement le modèle de Zurek. Le modèle de Zurek est conçu pour traiter le cas d'une source cible et d'une seule source masquante. Nous avons décidé d'extrapoler le principe au cas de N sources masquantes en considérant la somme des signaux masquants à chaque oreille.

On charge les C_i tels qu'ils ont été décris dans la partie 4.1.2. Ils ont été mesuré sur [38], mais les fréquences centrales des filtres utilisés par Zurek ne correspondent pas exactement à celles que nous avons utilisées. Etant donné l'allure de la courbe de C_i, il était cependant facile de les interpoler. Nous n'avons plus qu'à utiliser le calcul du MLD tel qu'il est décrit par Zurek, en applicant tout de même les modification expliquées dans le paragraphe précédent.

En dernière étape de la fonction `masking`, on somme le TMR obtenu en sortie de la fonction `TMR` avec le MLD puis on recale le TMR de façon à ce qu'il soit utilisable pour l'indice d'articulation :

$$TMR(instantN) = \begin{cases} 0 & TMR < -12 \\ TMR + 12 & -12 < TMR < 18 \\ 30 & TMR > 18 \end{cases} \tag{5.11}$$

Une dernière fonction `collecte` permet de collecter et de traiter les résultats de la fonction `masking`. C'est ici que l'idée de prédiction est appliquée sur le critère de qualité. Pour cela, on crée une fenêtre de prédiction de longueur L_{fen_predic}. Elle s'applique selon le même principe qu'une fenêtre d'oubli, mais regarde non pas les calculs du critère passés mais à venir (en partant du principe que l'on dispose déjà des évaluations du critère dans leur intégralité). On considère en effet qu'il est peu utile d'indiquer à l'utilisateur que la position actuelle de ses sources est inadaptée aux signaux qui ont été joués mais qu'il vaut mieux lui dire qu'elle est inadaptée (ou adaptée, selon les cas) aux signaux qui vont être joués dans les prochaines secondes. On a donc choisit de prendre une fenêtre exponentielle de la forme $e^{-\alpha n}$, donnant ainsi plus d'importance à ce qui est joué au moment où on évalue le facteur perceptif et de moins en moins d'importance pour ce qui sera joué dans un

futur de plus en plus éloigné.

$$\text{fen\_predic}(n) = \begin{cases} e^{\frac{-10n}{L_{fen\_predic}}} \text{ normalisé} & \text{si n est compris entre 0 et } L_{fen\_predic} - 1 \\ 0 & sinon \end{cases} \quad (5.12)$$

Le TMR_{final} est obtenu en intégrant $TMR(n)$ x $fen\_predic(n)$ entre l'instant courant n et l'instant $n + L_{fen\_predic} - 1$

La fonction `collecte` joue ensuite le vecteur de signal binaural obtenu et représente l'évolution du TMR en fonction du temps, soit sous forme de graphique, soit -et c'est ce qui nous intéresse ici- sous forme de niveaux de gris.

Nous avons également réalisé un programme pour traiter le cas stéréo et nous avons pu le tester avec succès.

5.3 Fonctions version stéréo

Nous ne présentons dans cette partie que les modifications effectuées aux fonctions décrites dans la partie 5.2 ainsi que les nouvelles fonctions.

5.3.1 Stereo et stereodB

Tout d'abord, nous avons écrit deux fonctions `stereo` et `stereodB`. Chacune de ces deux fonctions prennent en variable d'entrée une source ($s(n)$ signal temporel dans le cas de la première, et les $Y_i(n)$ pour la seconde), une distance d et un azimut az. Elles renvoient en sortie le contenu des enceintes gauche et droite si la source est placée à une distance d et un azimut az d'un couple de microphones XY. Dans le cas de `stereo`, il s'agit des signaux reçus par chacun des microphones virtuels $left$ et $right$(et donc envoyés à chacune des deux enceintes du système stéréo) que la fonction envoie en sortie. Dans le cas de `stereodB`, c'est les moyennes temporelles de chaque canal ERB $Left_i$ et $Right_i$ qui sont renvoyées en sortie.

$$left(n) = \frac{s(n)\ panleft(az)}{d} \quad (5.13)$$

$$right(n) = \frac{s(n)\ panright(az}{d} \quad (5.14)$$

avec

$$panleft(az) = (\sqrt{2} - 1)(1 + cos(\frac{\Pi}{4} - az)) \quad (5.15)$$

$$panright(az) = (\sqrt{2} - 1)(1 + cos(\frac{\Pi}{4} + az)) \quad (5.16)$$

dans le cas de `stereo`. Dans le cas de `stereodB`, il s'agit des même calculs mais dans un domaine logarithmique

5.3.2 masking

Les calculs du $Y_{i,left}$ et du $Y_{i,right}$ sont modifiés : pour chaque source, la fonction calcule d'abord les moyennes $Left_i$ et $Right_i$ issues de chaque enceinte à l'aide de la fonction `stereodB` puis celles de chaque oreille pour la source donnée. Le signal reçu à l'oreille gauche est la somme de celui envoyé par l'enceinte gauche pondéré par la $meanhrir_{left,i,30}$ et de celui envoyé par l'enceinte droite pondéré par la $meanhrir_{left,i,-30}$.

$$Y_{i,left}(N) = 20\log(Y1_{i,left} + Y2_{i,left}) \tag{5.17}$$

avec

$$Y1_{i,left} = 10^{(Left_i + meanhrir_{left,i,30})/(20)} \tag{5.18}$$

$$Y2_{i,left} = 10^{(Right_i + meanhrir_{left,i,-30})/(20)} \tag{5.19}$$

On calcule également non pas le signal binaural issu du placement de la source, mais le signal stéréo résultant à l'aide de la fonction `stereo`.

Le TMR est ensuite calculé de la même façon que dans la partie 5.2. Le calcul du MLD a cependant dû être modifié. En effet, on ne doit plus tenir compte pour le MLD de la source virtuelle mais des deux sources réelles que sont les haut-parleurs. Le calcul de la différence de niveau entre les deux oreilles reste le même. Il est toujours calculé à partir de $Y_{i,left}$ et $Y_{i,right}$ pour chaque source. Mais le calcul de la différence de phase est plus délicat : la différence de phase de l'enceinte gauche est l'opposé de la différence de phase de l'enceinte droite. Pour cela, on calcule une différence de phase moyenne à l'aide de pondérations : soit $\phi 1_i$ la différence de phase entre les oreilles due à l'enceinte gauche et $\phi 2_i$ la différence de phase entre les oreilles due à l'enceinte droite. soit α le coefficient de pondération : $alpha = 10.^{((Y_{i,left} - Y_{i,right})/20)}$. On a alors

$$\phi_i = \frac{\alpha\ \phi 1_i}{\alpha + 1} + \frac{\phi 2_i}{\alpha + 1} \tag{5.20}$$

Une erreur a été réalisée en faisant cela : cette mesure d'ITD n'a pas de valeur perceptive puisqu'elle ne correspond pas à un ITD réel. Il serait nécessaire d'utiliser des outils d'estimation de l'ITD par Inter-Aural Cross Correlation (IACC) en fonction du signal délivré par les enceintes. Cette estimation seule pourrait ne pas être suffisante (certains cas amèneraient trois pics de corrélation de même intensité), mais la méthode actuelle n'a pas de validité perceptive.

5.4 Risque d'être masqué : approche locale ou globale

La représentation précédente consiste à indiquer dans le fond de l'écran si une source sera ou non masquée si elle est placée en un certain point de l'espace. Cette représentation ne fournit pas toutes les informations nécessaires à l'utilisateur. En effet, ce dernier souhaitera généralement connaître le risque que n'importe quelle source de sa scène sonore soit masquée. Avec la méthode actuelle, il ne peut visualiser que le masquage d'une seule source par toutes les autres. Pour remédier à cela, nous avons décidé de conserver une source cible. On calcule en chaque point (x,y) de l'espace de visualisation le critère de masquage pour chacune des sources en supposant que la source cible se trouve en (x,y). On obtient un facteur de qualité Q pour chacune des sources, compris entre 0 (noir, on n'entend rien) et 1 (blanc, on entend parfaitement). On choisit de représenter pire d'entre eux.

$$Q_{reprsent,(x,y)} = \min_{i=1..nombredesources}(Q_{i,(x,y)}) \tag{5.21}$$

où $Q_{i,(x,y)}$ est la mesure du facteur de qualité pour la source i sachant que la source cible se trouve en (x,y) .

Il peut cependant arriver que quelle que soit la position de la source cible, le facteur de qualité d'une des sources soit toujours faible. La conséquence d'une telle situation est que le fond d'écran serait noir tout entier sans que l'utilisateur puisse savoir d'où vient le problème. La solution serait alors de calculer les facteurs de qualité pour chacune des sources sans tenir compte de la source cible. Si ce facteur de qualité est faible pour une des sources, cela signifie que peu importe la position de la source cible, le critère de masquage global mesuré sera toujours faible. Le problème est de représenter ces quelques facteurs de qualité sur l'interface de ListenSpace. Pour cela, on peut modifier la couleur de la source, passant d'un rose clair si le facteur de qualité (lorsque la source cible n'est pas là) à un rouge foncé si ce facteur de qualité est faible. Ainsi, si la visualisation du masquage global remplit tout le fond d'écran de noir, l'utilisateur saura quelle source est responsable de cela. La visualisation est alors légèrement moins intuitive mais a le mérite de tenir compte des facteurs de qualité de toutes les sources à la fois.

5.5 Implémentation sous ListenSpace

Une implémentation des algorithmes binaural et stéréo (couple XY) a été réalisée sous ListenSpace. Elle a été écrite en Java, comme le reste de ListenSpace. Cette implémentation permet d'évaluer en temps réel le risque qu'une source sonore soit masquée par les autres. Il est possible de paramétrer les position des sources, le système de restitution, la fenêtre de prédiction

et le modèle de spatialisation. Les métadonnées que ListenSpace reçoit ne le renseignent que sur les sources brutes : il s'agit de l'énergie de chacune des sources dans chacune des quinze bandes de fréquence des nos filtres ERB.

Chacune des deux classes commence par une étape d'instanciation des variables globales. Il s'agit de la plupart des variables dont nous avons besoin dans les méthodes utilisées. Parmi ces variables globales, on peut trouver les valeurs moyennes de chacune des HRIRs dans chaque bande fréquentielle. Plusieurs méthodes ont dû être redéfinies afin d'optimiser le temps de rendu du programme : une méthode de calcul de l'angle de la source ainsi que des méthodes de moyennage des HRTFs pour obtenir un résultat plus homogène. Il a également été nécessaire de redéfinir des méthodes de calcul matriciel, à l'image du fonctionnement de Matlab... Cela a permi un gain notable de rapidité de calcul. Nous avons également instauré une pondération, les résultats que nous obtenions ne nous semblant pas suffisamment réalistes sans cette pondération. Il s'agit de la pondération conseillée pour le calcul de l'indice d'articulation, comme indiqué dans [16]. Le nombre de paramètres qu'il est possible de contrôler dans ce modèle est élevé, même si par souci de simplicité, tout n'est pas directement accessible à l'utilisateur : système de restitution, choix de la base de HRTFs, position des sources, atténuation en distance respectant le modèle de spatialisation et prise en compte du champ diffus.

Nous avons rajouté dans le modèle de ListenSpace une certaine prise en compte du champ diffus. En effet, l'absence de modélisation du champ diffus au sein de l'algorithme pouvait fausser l'évaluation du critère de qualité. Lorsqu'un son est émis dans une salle, les ondes acoustiques se réfléchissent sur les murs et créent alors des sources fantômes qui émettent simultanément le même contenu que la source originale[1], mais qui ne sont pas situées à la même distance d'un point d'écoute. L'auditeur entend donc les différentes sources fantômes les unes après les autres[2]. La salle agit comme un filtre dont il est possible de mesurer la réponse impulsionnelle. Cette réponse impulsionnelle est souvent séparée en trois parties : le son direct, correspondant à la première impulsion de la réponse, les premières réflexions, correspondant généralement aux impulsions atténuées de moins de 10 dB par rapport au son direct, et enfin le champ diffus, correspondant au reste de la réponse impulsionnelle. Si il était délicat dans notre modèle de simuler les premières réflexions, nous avons simplement modélisé le champ diffus. Pour cela, ListenSpace reçoit de MAX/MSP et du Spat le niveau de champ diffus s_{diff} et celui du champ direct s_{direct} correspondant à la position de la source. On considère alors que le niveau de signal d'une source s reçu dans la bande i pour l'oreille gauche est

$$L_i = s_i \quad s_{direct,i} \quad gainhrir_{left,i,az_{source}} + s_i \quad * \quad s_{diff,i} \quad * \quad meanhrir \tag{5.22}$$

[1] atténué par l'absorption des parois

[2] même si il ne peut pas les dissocier

De même, pour l'oreille droite, on obtient

$$R_i = s_i \quad s_{direct,i} \quad gainhrir_{right,i,az_{source}} + s_i \quad * \quad s_{diff,i} \quad * \quad meanhrir \tag{5.23}$$

Si la source s est une source cible, on remplace s_{direct} par $\frac{s_{direct} dist_{source}}{dist_{pixel}}$, puisqu'on ne calcule pas le niveau de la source cible à sa position actuelle mais à la position du pixel. Le champ diffus est pondéré par une hrir moyenne car il est isotrope (et vient donc de toutes les directions).

Ce modèle du champ réverbéré n'est cependant pas parfait : il ne tient pas compte des premières réflexions (qui ne sont pas isotropes), et le modèle de champ diffus est ici le son direct pondéré par un facteur de champ diffus. Le champ diffus tel qu'il est modélisé par le Spat n'est plus corrélé avec le son direct...

Chapitre 6

Tests perceptifs

Suite à la description des implémentations d'algorithmes dans la partie précédente, ce chapitre expose les tests envisagés pour la validation des modèles de masquage. Il s'agit de tests similaires à des mesures de seuil de l'audition ainsi que d'un tests afin d'évaluer un mapping optimal des niveaux de gris vis à vis du TMR

Plusieurs séries de tests ont été réalisés : des tests préliminaires uniquement sous Matlab, des tests de validation de modèles sous MAX/MSP uniquement et enfin des tests de validation de l'implémentation des algorithmes sous ListenSpace.

6.1 Tests sous Matlab

Au cours de ce stage, nous avons commencé par implémenter les différents algorithmes que nous souhaitions tester sous Matlab. C'était en effet une solution simple et adaptée au traitement du signal. Cela permettait une vérification rapide de l'algorithme. Si au début du stage, une interface graphique avait été réalisée afin de vérifier les résultats, elle s'est avérée lourde à manipuler lorsque des modifications de l'algorithme devaient être faites. Nous avons donc conservé l'interface de Matlab et écrit un jeu de fonctions (leur version finale est décrite dans la section 5) que l'on modifiait en fonction de notre algorithme. Une de ces fonctions nous permettait de spécifier une scène sonore (quelles sont les sources et leurs positions en distance et azimut), d'appliquer l'algorithme dessus, d'écouter la scène sonore sur le système de restitution choisi pour l'algorithme et de comparer ce qu'on entendait aux résultats prédits par l'algorithme. Pour ces scènes sonores, nous disposions de trois sources. Les sources que nous avon utilisées étaient extraites d'un enregistrement multipistes de *Every breath you take*, de Police. Nous avons

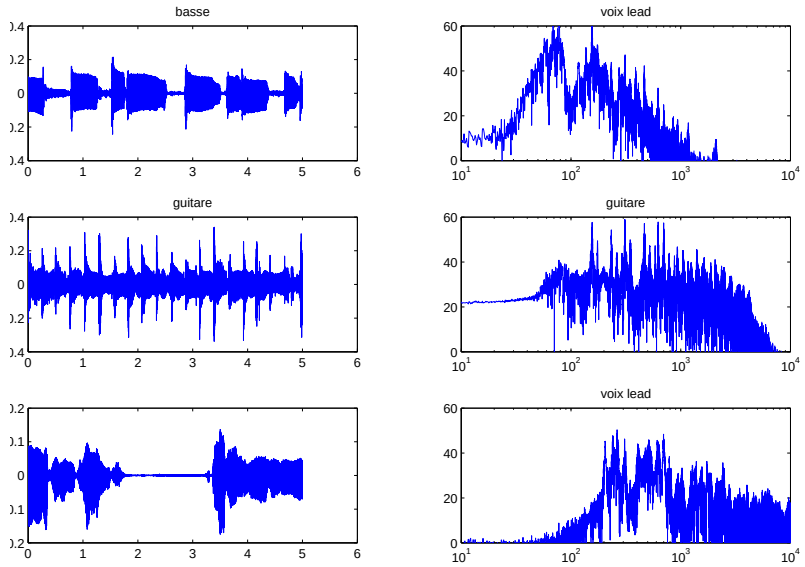

FIG. 6.1 – Contenu fréquentiel (à gauche) et temporel (à droite) des signaux utilisés. On note que la voix se tait au milieu de l'extrait

utilisé une piste de basse, la piste de voix lead ainsi qu'une piste de guitare électrique (voir fig. 6.1).

Il est à noter que si les pistes de basse et de guitare jouent approximativement tout le temps, la voix lead, elle, s'arrête quelques secondes. Nous étions libres de choisir quelle source serait la source cible et quelles sources seraient masquantes. Nous pouvions également choisir de n'utiliser que deux sources sonores. Si initialement, on pouvait choisir d'en utiliser plus que trois, cela s'est avéré inutile et nous nous sommes donc contentés, pour la plupart des tests, de trois sources.

Nous nous sommes intéressés dès le début à la méthode de Zurek. C'est en effet celle qui semble être unanimement choisi pour les études de masquage ou de démasqage spatial. En revanche, si l'idée générale était celle de Zurek, l'implémentation que nous avions testée initialement différait grandement dans son traitement. Plutôt que d'utiliser des filtres ERB, nous utilisions à l'origine une transformée de Fourier à court terme. Cette transformée était calculée de façon à ne pas avoir de problème de convolution circulaire lors de la multiplication des tranches fréquentielles avec les HRTFs.

On calculait ensuite le TMR de la manière suivante :

$$TMR_{left}(f_n, t_m) = \frac{|TFCTcible_{left}(f_n, t_m)| + \epsilon}{|TFCTmasker_{left}(f_n, t_m)| + \epsilon} \tag{6.1}$$

$$TMR_{right}(f_n, t_m) = \frac{|TFCTcible_{right}(f_n, t_m)| + \epsilon}{|TFCTmasker_{right}(f_n, t_m)| + \epsilon} \tag{6.2}$$

avec $f_n = \frac{n}{nfft\ Te}$ et t_m = numéro de la fenêtre. nfft est le nombre de points utilisés pour réaliser les fft. ϵ est un facteur qui permet d'éviter les problèmes d'inversion lors des divisions. En effet, nous prenions le module de la TFCT pour nos calculs, qui est toujours positif ou nul. Si on divise par une TFCT, on risque donc de devoir diviser par zéro. Pour éviter cela, on rajoute un epsilon au dénominateur. Il s'agit d'un nombre non nul mais très proche de zéro (non nul pour que le calcul soit réalisable, et très proche de zéro pour qu'il ne fausse pas les calculs). Nous nous sommes cependant aperçu que dans certaines conditions, bien que les calculs soient toujours réalisables, certaines fréquences du TMR présentaient de fort pics, même quand nous représentions le TMR en échelle logarithmique. Ces pics n'étaient cependant pas audibles. Ils pouvaient être dûs à l'absence d'énergie à certaines fréquences soit dans le signal cible, soit dans la combinaison des signaux masqueurs. Ce phénomène arrivait assez fréquemment. Nous avons donc décidé d'augmenter la valeur de cet epsilon. Il faussait alors les résultats. Son influence était moindre lorsque qu'on le rajoutait également au numérateur de la formule du TMR, mais les résultats n'étaient de toute façon pas fiables. Si l'allure des résultats obtenus était cohérente, les bornes de seuils de masquage et de niveau à partir duquel on entendait parfaitement une source étaient très différentes d'un signal à un autre. Nous sommes donc passés dans un premier temps à la méthode de Zurek telle qu'elle était décrite initialement puis l'avons modifié de manière à tenir compte du temps et de plusieurs sources et avons constaté que les résultats étaient beaucoup plus cohérents et que le seuil de perception ne dépendait plus des niveaux utilisés. Nous nous sommes également aperçu que lorsque l'on utilisait une source cible et seulement une source masquante, les résultats étaient similaires à ceux que l'on pouvait trouver dans la littérature sur le démasquage spatial, à savoir que le fait de déplacer une source cible d'une source masquante en azimut améliore notablement la perception de la source cible (voir fig. 6.2). Mais lorsque plusieurs sources masquantes sont présentes (par exemple : deux sources masquantes), l'intérêt du déplacement en azimut de la source cible est bien plus négligeable (voir fig. 6.3). On peut alors se poser des questions sur l'utilité d'un algorithme complexe pour représenter ce qui ressemblera au final à de simples anneaux de gris (dans le cas de plusieurs sources masquantes).

6.2 Tests sous MAX/MSP

Nous avons au cours de ce stage réalisé quelques modifications sur le modèle de Zurek tel qu'il était présenté à l'origine. Avant d'implanter l'algorithme sous ListenSpace, nous devons donc valider ces modifications à l'aide de tests perceptifs. Nous avons choisi de les réaliser sous MAX/MSP, puis de

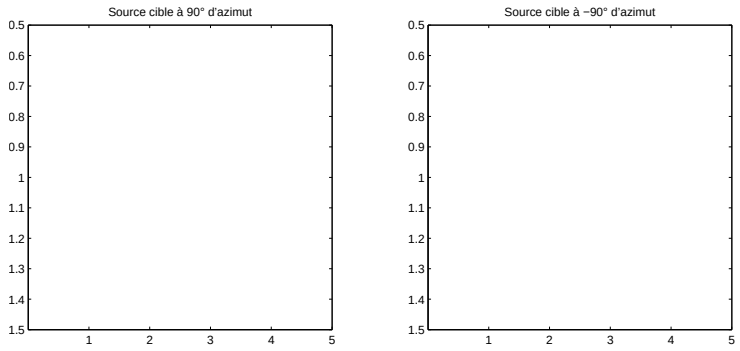

FIG. 6.2 – Courbes de masquage prédit lorsque la source cible est la guitare et qu'il n'y a qu'une source masquante : la voix. Elles sont placées respectivement à 10m, azimut 90°(à gauche) ou -90°(à droite) et 1m, azimut -90°

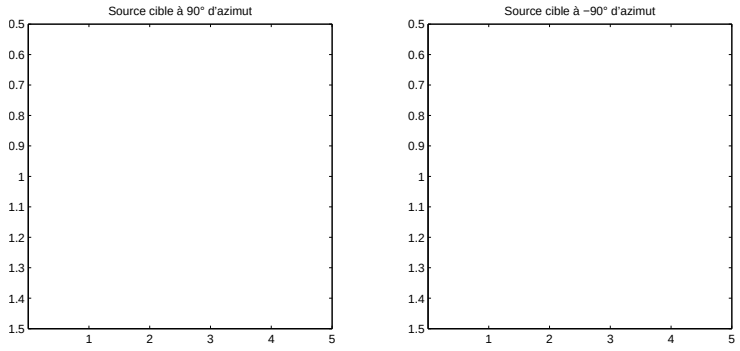

FIG. 6.3 – Courbe de masquage prédit lorsque la source cible est la guitare et qu'il y a les deux sources masquantes : la voix et la basse. Elles sont placées respectivement à 10m, azimut 90°(à gauche) ou -90°(à droite), 1m, azimut -90°et 1m, azimut 0°

les comparer aux prédictions de notre algorithme à l'aide de Matlab. Nous n'avons pas réalisé les expériences mais celles-ci sont prêtes et il nous faut procéder en deux étapes :

- Un test validant le seuil d'audibilité de notre modèle
- Un test validant le comportement angulaire du modèle ainsi qu'étudiant un mapping optimal des niveaux de gris vis-à-vis du TMR calculé.

6.2.1 Validation du seuil d'audibilité

Ce test devait ressembler à un test de mesure de seuil d'audition classique. En effet, nous voulions simplement étudier si dans une configuration donnée, le sujet était capable ou non d'entendre la source cible. Nous étions à même, dans cette configuration, de vérifier ce que prédisait le modèle de masquage.

Le programme est le suivant : Un patch (une fonction MAX/MSP) se charge des scènes sonores à l'aide d'un fichier texte créé préalablement sous Matlab. Ces scènes sonores comprennent une source cible (que le sujet doit reconnaître) et deux sources masquantes. Seule la source cible peut varier en distance, les deux sources masquantes étant placées à 1m du sujet. Une des deux sources masquantes est placée à l'azimut -90°, c'est à dire à droite du sujet. L'autre source masquante peut être placée à 0, -45 ou -90 degrés (respectivement devant, devant à droite et à droite du sujet). La source cible peut être placée à 90, 45, 0, -45 ou -90 degrés en azimut. La distance de la source cible varie de 1m à 26,6 m par pas d'un facteur 1,2 (afin que le pas en dB d'atténuation reste constant). Une scène consiste en une suite de nombre indiquant quelle est la distance de la source cible, les azimuts de la source cible et des deux sources masquantes ainsi que le contenu de la source cible. Ces scènes sont choisi au hasard. Il n'y a donc pas d'effet d'attente de la part du sujet, attente qui aurait été présente si on avait présenté une scène particulière et qu'on avait petit à petit éloigné la source cible comme sur un test de seuil de l'audition classique. On s'assure également que chaque scène est lue N fois, et on enregistre les résultats de ces occurrences.

Un autre patch, relié au premier, jouait simplement la scène sonore à l'aide des fichiers audio et de différentes instances du SPAT. Les signaux audio utilisés étaient de la parole. Les trois sources correspondaient à trois voix différentes. Le contenu des deux source masquantes était figé ("vacances" et "protéger") tandis que la source cible pouvait être soit "travailler", soit "fatigué", le contenu spectral et temporel de ces deux mots étant alors proche (voir fig. 6.4).

L'interface du sujet consiste en trois boutons. On demande à l'utilisateur si il a ou non entendu le mot "fatigué". Il dispose de trois boutons : un bouton "fatigué", non pas pour nous indiqué qu'il est fatigué et qu'il veut arrêter l'expérience mais pour dire au programme qu'il a entendu le mot fatigué. Un bouton "travailler" pour dire qu'il a entendu le mot "travailler"

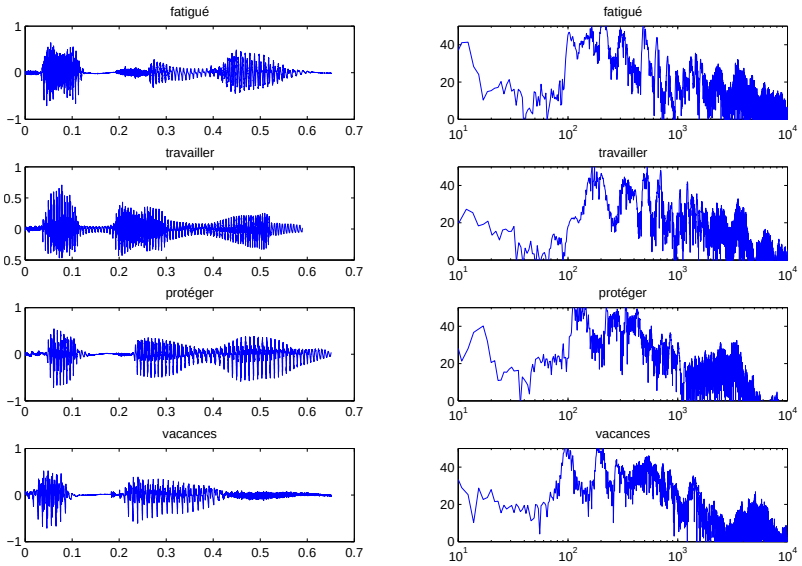

FIG. 6.4 – Représentation temporelle (à gauche) et fréquentielle (à droite) des différents mots prononcés.

et enfin "je ne sais pas" lorsqu'il ne sait pas ce qu'il a entendu.

On enregistre ensuite les résultats : correct si le sujet a reconnu le mot, et incorrect si il s'est trompé ou bien si il a appuyé sur le bouton "je ne sais pas". Les résultats sont enregistrés dans un fichier texte séparé, différent pour chaque sujet.

On calcule ensuite pour chaque configuration le pourcentage de bonnes réponses, en mélangeant les configurations où le mot prononcé était "travailler" et celles où le mot prononcé était "fatigué".

En reliant les résultats avec ceux du modèle implémenté sous Matlab, on s'intéresse à plusieurs éléments : A quel pourcentage de bons résultats correspond le seuil de perception prédit ? Un même TMR sur différentes configurations amène-t-il un même pourcentage de résultats corrects ?

Avant de commencer l'expérience en tant que tel, le sujet est invité à procéder à un phase d'entraînement au cours de laquelle un retour visuel lui indique si la réponse qu'il a donnée est correcte.

Mais cette expérience ne nous permet de valider que le seuil de perception du modèle. Pour une validation plus approfondie du modèle, il faudrait étudier la constance de l'impression perceptive pour différentes configurations spatiales qui donnent un même TMR.

Chapitre 7

Perspectives

A la suite des validations de notre modèle, ce chapitre s'intéresse aux perspectives de recherche sur l'aide au mixage et relève différents problèmes à résoudre / différentes pistes de recherche future.

Ce stage a soulevé un certain nombre de problèmes et de points délicats. Si le modèle de masquage utilisé a été avantageusement améliorer, de nombreuses améliorations sont possibles.

La représentation temps/fréquence utilisée actuellement est un banc de filtres destructif, dans le sens où il n'est pas possible de reconstituer le signal exact à partir des signaux filtrés. Il nous est actuellement impossible, ou au moins peu recommandé, de décimer les signaux issus des filtres. On multiplie donc la quantité d'information sans pour autant être capable de reconstituer le signal original. L'avantage de la méthode actuelle est la similitude entre les filtres utilisés et le comportement des cellules ciliaires. C'est probablement la raison de son utilisation fréquente dans les applications de détection des sons. Il serait néanmoins très intéressant de tester des bancs de filtres de type transformée en ondelettes ou d'autres types de bancs de filtres à largeur de bande variable.

L'idée de l'application des cartes de saillance au domaine de l'audio pourrait également être intéressante et permettre d'optimiser le modèle actuel. En effet, un point de vue du modèle que nous utilisons actuellement est que nous décrivons chaque signal comme une matrice temps/fréquence (plusieurs canaux fréquentiels et une évolution temporelle de signaux ou de leur moyenne), matrice sur laquelle nous appliquons des transformations visant à calculer un rapport signal/bruit particulier. Nous pourrions, avec l'aide des cartes de saillance, optimiser la représentation temps/fréquence de l'algorithme de prédiction et ainsi se rapprocher de notre perception des sons dans l'espace.

Une des modifications que nous avons apportés au modèle de Zurek nous

permet de tenir compte du temps et de la macro-évolution de ces signaux dans le temps. Les tests de validation que nous avons réalisé ne tenaient pas compte de cette évolution dans le temps. Il faudrait donc réaliser une expérience visant à valider l'implémentation temporelle que nous avons faite.

Dans le modèle actuel, nous ne tenons pas réellement compte de l'ITD (Interaural Time Difference). Cet ITD est tout de même utilisé dans la formule du MLD. De récents travaux ont fait part de l'importance de l'ITD dans la discrimination de sources dans une scène sonore. Nous devrions en tenir compte étudier comment l'ITD pôurrait améliorer notre modèle. En revanche, les tests que nous avons effectué nous montrent que lorsque plusieurs sources masquantes sont présentes, l'azimut influe moins sur le seuil de perception d'une source cible.

Comme indiqué dans la partie 4.1.3, il faudrait tenir compte (si le système de spatialisation en tient compte également) de la différence de chemin acoustique des sources à chacune des deux oreilles ainsi que du champ diffus. Ces améliorations sembleraient d'ailleurs plus importantes que les améliorations possibles du modèle à l'aide de l'ITD. Nous n'avons pas utilisé de pondération fréquentielle dans notre modèle, contrairement au modèle original de Zurek. En effet, cette pondération avait été mise en place dans le cadre de l'indice d'articulation de la voix. Nous ne souhaitions pas nous limiter à la voix mais à tout contenu musical. Nous avons donc préféré dans un premier temps nous affranchir de cette pondération. Il serait souhaitable de la réintégrer à l'algorithme afin de tester sa validité dans le cadre d'un contenu musical.

Dans le modèle de masquage actuellement implémenté, nous ne tenons pas compte de la corrélation entre les sources cible et masquantes. Dans certains cas particuliers, lorsque la corrélation entre plusieurs sources est forte, le masquage peut être influencé : si la source cible est fortement corrélé avec une source masquante, l'utilisateur pourra ne plus entendre la source cible mais une image fantôme qui apparaîtra entre elle et et la source masquante. Cela ne pose pas de problème à priori dans la mesure où la source fantôme est très semblable à la source cible. en revanche, si deux sources masquantes de part et d'autre d'une source cible sont fortement corrélées, la source fantôme résultante apparaîtra entre les deux sources masquantes, soit dans la même direction que la source cible. Cela peut nuire à la perception de la source cible.

Les tests perceptifs préparés sous MAX/MSP n'ont pas été utilisés sur des sujets. La raison à cela est que les sons à reconnaître étaient trop courts. Pour remédier à cela, il faudrait inclure les mots dans des phrases. Les trois phrases entendues doivent être différentes. Une seule des trois (précisée au début du test) contient le mot cible. Des tests proches de celui que nous avons voulu faire ont déjà été réalisés au cours d'un stage co-encadré par Olivier Warusfel et Alain De Cheveigné. Il nous a donc semblé préférable, plutôt que de mettre en place un tel test, de consacrer notre temps à étudier

d'autres critères...

Par extension de ceci, il serait intéressant de s'intéresser à la formation d'images fantômes. Cela permettrait à l'utilisateur de conserver une certaine fidélité de la scène sonore si il manipule plusieurs sources corrélées comme des sources issues de couples microphoniques ou d'extraction de scènes sonores. L'interface pourrait alors également mettre en garde l'utilisateur contre un détimbrage possible (par exemple lorsque deux sources corrélées ne sont pas à la même distance de l'utilisateur, elles arrivent à des temps différés aux oreilles de l'utilisateur, créant ainsi un filtrage en peigne).

On pourrait enfin s'intéresser à des considérations plus artistiques en étudiant comment est généralement construite une scène sonore par les ingénieurs du son. Mais cela s'éloigne plus de l'aspect traitement du signal que nous nous étions fixés.

Chapitre 8

Conclusion

Au cours de ce stage, nous avons étudié différents critères de qualité basés sur des facteurs perceptifs dans le but de les représenter au sein d'une interface de visualisation et de manipulation de scènes sonores.

Pour cela, nous nous sommes intéressés aux systèmes de restitution, leur influence étant très importante dans le calcul des critères de qualité puisque les signaux qui arrivent à chacune de nos deux oreilles dépend de la configuration de haut parleurs utilisée.

Nous nous sommes intéressés principalement à trois facteurs perceptifs : le masquage spatial, la perception d'artefacts lors du mixage de sources issues de processus de séparation de sources et la formation d'images fantômes. La majeure partie de nos travaux a porté sur le premier, en s'appuyant sur un modèle déjà existant et en améliorant certains points (prise en compte de plusieurs sources masquantes et du champ diffus). Nous avons également constaté la limitation de ce modèle à la parole. En effet, il semble bien moins adapté à l'écoute de musique : il étudiait d'intelligibilité et non l'audibilité d'un signal[1]. Ces limitations ont pu être mises en évidence lors de tests perceptifs informels réalisés via l'implémentation temps réel de ce modèle amélioré sous ListenSpace.

Nous avons également mis en évidence d'autres facteurs perceptifs intéressants tels que la largeur de sources sonores ou la perception d'artefacts lors de spatialisation de sources sonores démixées. Nous nous sommes intéressés à ce dernier point, mais n'avons à l'heure actuelle que peu d'éléments pour mettre en évidence cette perception d'artefacts. Pour ces différents problèmes, nous avons mis en évidence les indices objectifs importants pour la mise en place de critères de qualité adaptés, sans pour autant réaliser de modèles détaillés pour ces facteurs perceptifs.

Les prochaines étapes importantes de ce projet seront une prise en compte dans ListenSpace du modèle de champ diffus du Spat ainsi qu'une validation

[1] Le modèle de Zurek cherchait non pas savoir si on peut percevoir un signal audio mais si on peut le comprendre

perceptive du modèle. La perception d'artefacts devra également être étudiée plus en profondeur.

Bibliographie

[1] http ://shf.ircam.fr/.

[2] Jens Blauert. *Spatial Hearing*. The MIT press, 1997.

[3] Sylvain Busson. *Individualisation de la synthèse binaurale*. PhD thesis, Université de la mediterranée Aix-Marseille II, 2006.

[4] Emmanuel Vincent Cédric Févotte, Rémi Gribonval. Bss_eval toolbox user guide revision 2.0.

[5] R. I. Chernyak and N. A. Dubrovsky. Pattern of the noise images and the binaural summation of loudness for the different interaural correlation of noise. Proceedings, 6th Int. Congr. on Acoustics, vol.1, Tokyo, 1968.

[6] H. Steven Colburn. Theory of binaural interaction based on auditory-nerve data. i. general strategy and preliminary results on interaural discrimination. *J. Acoust. Soc. Am*, 54(6) :1458–1470, December 1973.

[7] Olivier Delerue. *Spatialisation et programmation par contraintes : le système MusicSpace*. PhD thesis, Université Paris 6, 2004.

[8] Olivier Delerue. Visualization of perceptual parameters in interactive user interfaces : Application to the control of sound spatialization. 120th Convention of the Audio Engineering Society, Paris, France, May 2006.

[9] Nathaniel I. Durlach, Christine R. Mason, Frederick J. Gallun, Barbara Shinn-Cunningham, H. Steven Colburn, and Gerald Kidd Jr. Informational masking for simultaneous nonspeech stimuli : Psychometric functions for fixed and randomly mixed maskers. *J. Acoust. Soc. Am*, 118(4) :2482–2497, October 2005.

[10] Cédric Févotte Emmanuel Vincent, Rémi Gribonval. Performance measurement in blind audio source separation. *IEEE*, 2004.

[11] Christof Faller and Juha Merimaa. Localization in complex listening situations : Selection of binaural cues based on interaural coherence. *J. Acoust. Soc. Am*, 116(5) :3075–3089, 2004.

[12] Richard L. Freyman, Karen S. Helfer, Daniel D. McCall, and Rachel K. Clifton. The role of perceived spatial separation in the unmasking of speech. *J. Acoust. Soc. Am*, 106(6) :3578–3588, December 1999.

[13] Frederick J. Gallun, Christine R. Mason, and Gerald Kidd Jr. Binaural release from informational masking in a speech identification task. *J. Acoust. Soc. Am*, 118(3) :1614–1625, September 2005.

[14] Monica L. Hawleyb, Ruth Y. Litovskyc, and John F. Culling. The benefit of binaural hearing in a cocktail party : Effect of location and type of interferer. *J. Acoust. Soc. Am*, 115(2) :833–843, February 2004.

[15] Toni Hirvonen and Ville Pulkki. Center and spatial extent of auditory events as caused by multiple sound sources in frequency-dependent directions. *Acta Acustica*, in press, 2006.

[16] American National Standard Institute, 1969.

[17] L. A. Jeffress, H. C. Blodgett, and B. H. Deatherage. The masking of tones by white noise as a function of the interaural phases of both components. *J. Acoust. Soc. Am*, 24 :523–527, 1952.

[18] Jean-Marc Jot. *Etude et Realisation d'un spatialisateur de sons par odèles physiques et perceptifs*. PhD thesis, Telecom Paris, 1992.

[19] Jean-Marc Jot. Efficient models for distance and reverberation rendering in computer music and virtual audio reality. Proceedings of the International Computer Music Conference ICMC97, 1997.

[20] Gerald Kidd Jr., Christine R. Mason, and Frederick J. Gallun. Combining energetic and informational masking for speech identification. *J. Acoust. Soc. Am*, 118(2) :982–992, August 2005.

[21] Alan Kan and André van Schaik. Psychoacoustic evaluation of a new method for simulating near-field virtual auditory space. Audio Engineering Society, 120th convention, May 2006.

[22] Christoph Kayser, Christopher I. Petkov, Michael Lippert, and Nikos K. Logothetis. Mechanisms for allocating auditory attention : An auditory saliency map. *Current Biology*, 15 :1943–1947, 2005.

[23] Karl D. Kryter. Methods for the calculation and use of the articulation index. *J. Acoust. Soc. Am*, 34(11) :1689–1397, November 1962.

[24] Brad Libbey and Peter H. Rogers. The effect of overlap-masking on binaural reverberant word intelligibility. *J. Acoust. Soc. Am*, 116(5) :3141–3151, November 2004.

[25] Eli Osman. A correlation model of binaural masking level differences. *J. Acoust. Soc. Am*, 50(6B) :1494–1495, December 1971.

[26] Munhum Park, Philip A. Nelson, and Youngtae Kim. An auditory process model for the evaluation of virtual acoustic imaging systems. Audio Engineering Society 120th Convention, Paris, 2006.

[27] Lord Rayleigh. On our perception of sound direction. *Phylosophy Magazine*, 13 :214–232, 1907.

[28] Sylvain Rousselle. S´eparation de la voix du locuteur et du fond musical dans des emissions radiodiffusées. Technical report, ENST, 2006.

[29] Jacob W. Scarpaci, N.I. Durlach, and H. Steven Colburn. Binaural vs. better-ear listening.

[30] K. D. Schenkel. Accumulation theory of binaural-masked thresholds. *J. Acoust. Soc. Am*, 41(1) :20–31, January 1967.

[31] Barbara G. Shinn-Cunningham. Speech intelligibility, spatial unmasking, and realism in reverberant spatial auditory displays, 2002.

[32] Barbara G. Shinn-Cunningham. Spatial hearing advantages in everyday environments, 2003.

[33] Barbara G. Shinn-Cunningham, Erol Ozmeral, Frederick J. Gallun, Kamal Sen, and Virginia Best. Spatial unmasking of birdsong in human listeners : Energetic and informational factors. *J. Acoust. Soc. Am*, 118(6) :3766–3773, December 2005.

[34] Barbara G. Shinn-Cunningham, Jason Schickler, Norbert Kop co, and Ruth Litovsky. Spatial unmasking of nearby speech sources in a simulated anechoic environment. *J. Acoust. Soc. Am*, 110(2) :1118–1129, August 2001.

[35] Barbara G. Shinn-Cunningham, Jason Schickler, Norbert Kopco, and Ruth Litovsky. Spatial unmasking of nearby speech sources in a simulated anechoic environment. *J. Acoust. Soc. Am*, 110(2) :1118–1129, August 2001.

[36] Richard M. Stern and Constantine Trahiotis. Models of binaural perception. 1996.

[37] M. Williams. Unified theory of microphone systems for stereophonic sound recording. Audio Engineering Society 82nd convention, February 1987.

[38] Patrick M. Zurek. *Acoustical factors affecting hearing aid performance*, chapter Binaural advantages and directional effects in speech intelligibility. Allyn and Bacon, Boston, ii edition, 1993.

[39] Patrick M. Zurek, Richard L. Freyman, and Uma Balakrishnan. Auditory target detection in reverberation. *J. Acoust. Soc. Am*, 115(4) :1609–1620, April 2003.