THÈSE / UNIVERSITÉ DE BRETAGNE OCCIDENTALE

sous le sceau de l'Université Bretagne Loire

pour obtenir le titre de

DOCTEUR DE L'UNIVERSITÉ DE BRETAGNE OCCIDENTALE

Mention :

École Doctorale SICMA

présentée par

Andrey Fedosov

Préparée à b-com, Rennes

Sciences et Technologies de l'Information et de la Communication

Assistance automatique au mixage de microphones d'appoint dans une prise de son HOA. Etude exploratoire

Thèse soutenue le 15/02/2017

devant le jury composé de :

Sylvain Marchand
Professeur, Université de La Rochelle (La Rochelle) / *directeur de thèse*

Grégory Pallone
Ingénieur R&D, Orange (Cesson-Sévigné) / *encadrant*

Jérôme Daniel
Ingénieur R&D, Orange (Lannion) / *encadrant*

Frédéric Bimbot
Directeur de Recherche Inria. Irisa (Rennes) / *rapporteur*

Gaël Richard
Professeur, Télécom ParisTech (Paris) / *rapporteur*

Gilles Coppin
Professeur, Télécom Bretagne (Brest) / *examinateur*

Richard Kronland-Martinet
Directeur de Recherche, Directeur de Recherche CNRS au LMA (Marseille) / *examinateur*

Contents

Abstract

In this thesis we study the problematic of a sound engineer mixing HOA (Higher Order Ambisonics) and spot microphones, namely the estimation of parameters such as delay, position and gain of acoustic sources associated to spot microphones. We present a typical workflow in this context, and also propose an algorithm extracting parameters that could be applied to the spot microphone signals. This mixing assistance allows sound engineers to easily work with HOA 3D sound and to concentrate on artistic choices (fine adjustments of the parameters), by avoiding a low-added value work (coarse parameter estimation). The robustness of the estimators is evaluated on recorded and artificial sound scenes, with different degrees of complexity in terms of number of sources and acoustic conditions (reverberation, effect of real microphone encoding, …). We also provide performance evaluations, based on both sound scene simulations and real recordings, showing encouraging results along with actual limits, and conclude on perspectives.

Résumé

Dans ce travail nous étudions la problématique des ingénieurs du son face au mixage d'un microphone principal HOA avec des microphones d'appoint, et notamment l'estimation des paramètres tels que le retard, la position et le gain des sources acoustiques associées aux microphones d'appoint. Nous proposons un algorithme fournissant les paramètres estimés (retard, position, gain) basé sur des équations d'encodage spatial au format HOA qui peuvent alors être utilisées pour traiter les signaux des microphones d'appoint durant le mixage. Cette extraction automatique des paramètres peut être vue comme une assistance pour les ingénieurs du son, leur permettant d'éviter un travail à faible valeur ajoutée (mesure de la distance et des angles entre microphones) afin de pouvoir se concentrer sur des problèmes artistiques comme l'ajustement des paramètres de niveau, d'égalisation ou de compression, voire l'ajustement fin des paramètres de retard, position, gain. La robustesse de l'algorithme est présentée pour les scènes sonores de différents niveaux de complexité (plusieurs sources acoustiques, réverbération, encodage réel du microphone…). Nous proposons des tests de performances pour les scènes sonores simulées et réelles afin de montrer l'efficacité de l'algorithme ainsi que ses limites. La conclusion et les perspectives pour des futurs travaux complètent cette thèse à la fin du document.

1 Introduction

1.1 Cadre de la thèse

Ce travail vise globalement à démocratiser le format audio 3D prometteur HOA (« Higher Order Ambisonics ») à travers la mise à disposition des ingénieurs du son d'outils leur simplifiant son usage.

Grâce à l'intérêt grandissant dans le son 3D et la disponibilité des microphones 3D, mais aussi en considérant les avantages (simplicité d'enregistrement, indépendance au système de prise de son et de restitution) de la technologie HOA et son support dans le nouveau codec MPEG-H 3D Audio, les enregistrements HOA vont probablement gagner en popularité durant la prochaine décennie.

Dans ce contexte, nous étudions la problématique des ingénieurs du son face au mixage d'un microphone principal HOA avec des microphones d'appoint, et notamment l'estimation des paramètres tels que le retard, la position et le gain des sources acoustiques associées aux microphones d'appoint. Cette approche du mixage est actuellement largement répandue en stéréo dans la musique classique, mais elle pourrait être également utilisée dans d'autres applications professionnelles (théâtre et cinéma), mais aussi dans des applications grand public (répétitions, enregistrement d'évènements familiaux).

Les travaux effectués font partie de la recherche industrielle au sein de l'institut de recherche technologique « b-com ». Les briques technologiques présentées dans cette thèse sont à l'origine de la publication de l'article [1] et du dépôt de brevet [2], et liées à la production d'un plugin audio de DAW (« Digital Audio Workstation ») dont l'algorithme a été prototypé en Matlab.

1.2 Organisation du document

Le document est divisé en 6 chapitres représentant le sujet principal de ce travail. Dans le chapitre 2 nous étudions la chaîne de production audio dans un cas classique (stéréo) et introduisons le terme « mixage » ainsi que « microphone principal » et « microphone d'appoint ». Le passage de la stéréophonie vers la technologie « ambisonics » montre l'évolution des systèmes de prise de son. Nous décrivons la technique de prise de son classique du domaine stéréo. Nous présentons ensuite deux types d'appareils microphoniques du domaine ambisonique : au format B (Ambisonics d'ordre 1) et aux ordres supérieures HOA (« Higher

Order Ambisonics »). Les prises de son effectuées dans le cadre de cette thèse sont également présentées dans cette section du travail. Nous évoquons le principe du mixage et les paramètres importants à ajuster avant de les appliquer pendant le mixage ainsi que les difficultés dans le cas d'une mauvaise estimation des paramètres. A la fin du chapitre 2 la chaîne de production audio immersif et le module « Assistance automatique au mixage HOA » relié à l'estimation de paramètres du mixage (retard, azimut, élévation et gain) sont étudiés dans la conception de la problématique globale.

Dans le chapitre 3 nous présentons les méthodes d'estimation des paramètres du mixage en deux parties : pour le domaine temporel et fréquentiel. L'état de l'art est fait pour chaque domaine afin de trouver les méthodes qui peuvent être potentiellement proches de celles proposées dans ce travail. La fonction d'intercorrélation en tant qu'outil principal fait l'objet d'études particulières pour estimer le retard. Nous proposons une nouvelle fonction d'intercorrélation qui améliore les résultats d'estimation et fait l'objet d'une demande de brevet [2]. L'estimation de la position et du gain qui est basée sur les premières 4 composantes HOA sont ensuite étudiée. Dans ce chapitre nous introduisons la notion de « descripteur » et « d'indice de confiance » pour pouvoir juger l'estimation effectuée et sélectionner des paramètres bien estimés. Cette contribution fait partie de l'extension du brevet [2].

A la base des méthodes d'estimations proposées nous montrons dans le chapitre 4 l'algorithme d'estimation avec les schéma-blocs ainsi que la structure d'estimation : segmentation du signal par les « blocs » et « trames ». Plusieurs schémas seront proposés (domaine temporel et fréquentiel, estimation « bloc par bloc » et « trame par trame ») afin de pouvoir intégrer l'algorithme dans un plugin audio de DAW.

Dans le chapitre 5 nous proposons des tests de l'algorithme à partir d'un code réalisé en Matlab. Les tests sont composés de scènes sonores simulées (à partir des fichiers audio on construit la scène sonore au format HOA, introduit le délai dans certains signaux, ajoute la réverbération, etc.) et d'une scène sonore réelle enregistrée par le microphone ambisonique « Eigenmike ». A la fin de chaque test une conclusion est donnée afin de valider l'efficacité de l'algorithme.

Nous faisons la conclusion globale dans le chapitre 6 qui rappelle les résultats importants obtenus dans ce travail ainsi que les perspectives et les futurs travaux possibles dans ce domaine de recherche.

2 La production de contenus « audio immersif »

Le terme « mixage » ou simplement « mix » désigne un ensemble d'opérations de traitement de signaux audio, réalisées par un logiciel ou par un appareil, au terme desquelles tous les signaux sont mélangés pour obtenir un son unifié. Ce mélange est effectué en réglant le niveau sonore, le timbre, la spatialisation et d'autres caractéristiques sonores des sources. En général, ce son est constitué de plusieurs signaux distincts, diffusés sur plusieurs haut-parleurs distribués dans l'espace d'écoute (ou au casque), afin de créer une image de scène sonore où l'on peut percevoir des sources localisées en azimut, élévation et profondeur (c'est la « stéréophonie », au sens large). L'étape de « mixage », réalisée par exemple dans un studio d'enregistrement, est une partie de la production audio : musique, films, émissions radio et TV.

Les étapes importantes de la production audio dans un cas classique sont la captation du son par des capteurs acoustiques, l'enregistrement dans un format spécial (par exemple fichier audio), le réglage (ajustement des paramètres reliés aux caractéristiques sonores du son traité) et le mixage des signaux traités avant les restituer vers un système de diffusion du son (Figure 1).

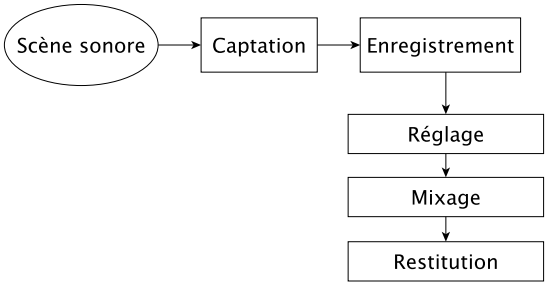

Figure 1. Production de contenus audio dans un cas classique.

Dans une conception habituelle (enregistrement de musique classique par exemple), la prise de son d'une scène sonore consiste en l'utilisation d'un système microphonique *principal* qui fournit une image sonore globale de la scène tout en apportant la couleur et le volume de l'espace. Bien souvent, chaque microphone du système capte un signal qui est ensuite restitué sur un haut-parleur dédié (approche multicanale). L'image sonore qui en résulte dépend des différences d'amplitude et/ou de phase entre un même signal diffusé par différents haut-parleurs. Pour définir plus précisément des sources acoustiques importantes, le preneur du son utilise des microphones d'*appoint*, disposés à proximité des sources en question (Figure 2).

Figure 2. Microphones d'appoint à proximité de chaque source acoustique potentielle dans l'Espace de Projection de l'IRCAM à Paris.

Avant de traiter les signaux captés par les microphones il faut les enregistrer, c'est-à-dire stocker l'information dans un format adéquat (par exemple mono ou stéréo, au format WAV, AIFF…). Au moment d'effectuer le mixage l'ingénieur du son ajuste des paramètres de mixage à partir des signaux captés en prenant compte certaines caractéristiques de la scène sonore comme la disposition des sources acoustiques, celle des microphones d'appoint, la présence de réverbération, etc. En particulier, pour ajuster les signaux sonores l'ingénieur du son doit connaître la distance entre les microphones d'appoint associés aux sources acoustiques et le microphone principal afin de compenser le temps de propagation des ondes acoustiques depuis les sources jusqu'au microphone principal. Un autre paramètre est la position angulaire des différentes sources sonores (ou leur azimut dans le cas où elles sont localisées dans le plan horizontal). L'estimation de la position des microphones d'appoint relativement au microphone principal se fait parfois par la mesure. Les erreurs de mesure, et le fait que la distance entre la source et le microphone d'appoint soit souvent négligée, font qu'il est nécessaire d'ajuster les paramètres à l'oreille pendant le mixage afin d'éviter des artefacts audio (filtrage en peigne, instabilité des sources, etc.).

Dans une conception du son immersif (Figure 3) la création d'un champ sonore à partir de l'enregistrement et du mixage réalisés prend en compte une autre caractéristique, dite « d'immersion sonore ». D'un point de vue perceptif l'immersion sonore pour un auditeur consiste en une sensation auditive et, par conséquent, une appréciation subjective, dans la scène sonore présentée par un système de diffusion du son (par exemple les haut-parleurs).

Une scène sonore composée de sources acoustiques est captée par un ou plusieurs microphones. L'étape de captation du son immersif est différente de celle dans la conception classique car elle consiste à utiliser des microphones spéciaux afin d'enregistrer le son dans un format spatial. Il existe trois approches du format spatial [3]: « channel-based » lorsque chaque canal est envoyé vers un haut-parleur dans une direction particulière (par exemple le système 5.1) ; « object-based » représente le son comme un élément séparé (par exemple « voix », « batterie ») accompagné d'une information sur position spatiale associée; « scene-based » basé sur les composantes spatiales du champ sonore, principe sur lequel repose cette thèse et qui sera adressé en détail plus loin.

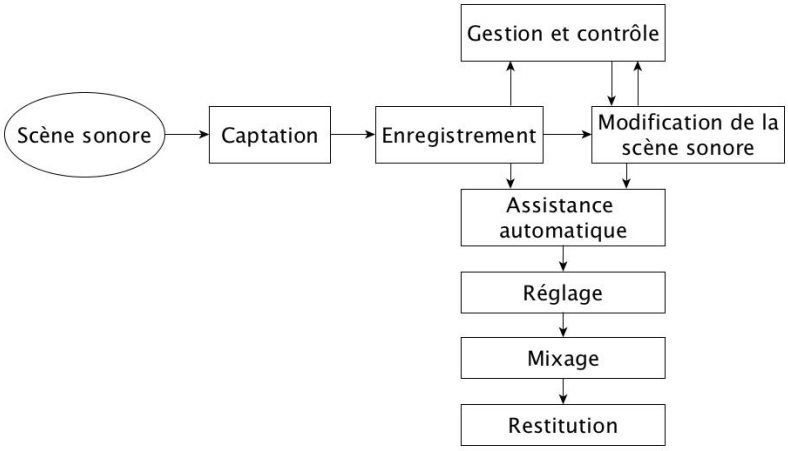

Figure 3. La chaîne de production audio immersif.

L'ingénieur du son possède à sa disposition les signaux du microphone principal et des microphones d'appoint. Avec ce format spatial l'ingénieur du son est capable d'effectuer le contrôle et des modifications de la scène sonore (par exemple, rotation de la scène en espace). La problématique du mixage du son immersif est proche de celle du mixage classique et reliée aux ajustement des sources acoustiques. Le but du travail de cette thèse est de proposer une assistance automatique pour fournir à l'ingénieur du son une estimation des paramètres d'ajustement qui lui seront utiles lors l'étape de mixage.

D'une part, l'assistance automatique permet d'éviter les travaux de mesures « à la main » (avec un mètre par exemple) ou « à l'oreille » (en agissant sur le retard du son) de la distance entre chaque source acoustique et le microphone principal. Dans le cas de sources acoustiques se déplaçant dans l'espace, l'assistance automatique devient quasiment indispensable car elle peut fournir les paramètres estimés au cours de temps. D'autre part, le mixage est une étape de post-production qui se déroule après l'enregistrement et prend un certain temps avant d'obtenir un résultat final. Grâce à l'assistance automatique l'ingénieur du son pourrait effectuer le mixage du son immersif avec les microphones d'appoint en temps réel

pendant une session d'enregistrement. Cependant l'assistance automatique n'effectue pas le mixage et ne fait pas le travail pour l'ingénieur du son car le mixage final reste toujours une étape à la fois technique, mais aussi et surtout artistique.

2.1 Prise de son stéréo et mixage classique

Une approche simple pour la captation du champ sonore immersive est la prise de son stéréo (« stéréo » vient du terme « stéréophonie » introduit en France en 1924 par l'ingénieur du son Georges-Clément Lévy : « Elle sera pour l'oreille ce que les yeux attendent du cinéma en relief »). La captation du champ sonore la plus usitée repose sur la prise de son à l'aide de couples microphoniques pour une restitution stéréophonique sur deux haut-parleurs, dont les principes remontent aux années 1930. Les techniques de prise de son stéréo se divisent en deux familles selon que les capsules des microphones sont coïncidentes ou non-coïncidentes (chaque technique est présentée en détails dans l'Annexe 1). Dans la technique coïncidente, les 2 capsules d'un couple stéréophonique sont placées extrêmement proches l'une de l'autre [4]. Un des exemple de ce système est une couple MS qui est composée d'un microphone (« M ») de type omni ou cardioïde placé en face de la source acoustique et d'un deuxième microphone (« S ») bidirectionnel dont la capsule est positionnée latéralement par rapport à la capsule « M » (Figure 4).

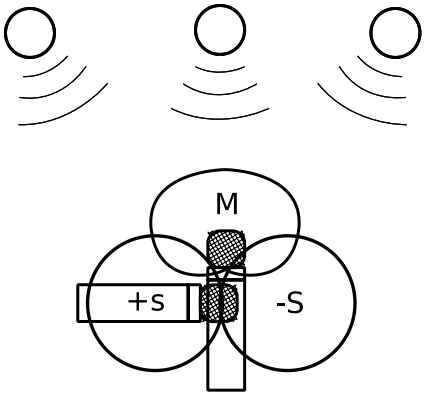

Figure 4. Représentation de la technique MS de captation du son par le microphone stéréo.

Dans le cas des techniques de prise de son non-coïncidentes les capsules sont espacées ce qui fournit un champ sonore moins précis mais plus ample [4]. Les travaux du pionnier de la stéréo, Alan Blumlein, ont établi les bases des systèmes modernes de prise de son et de restitution pour la TV et la Radio [5].

La prise de son par des microphones stéréo prend en compte des détails techniques tels que le placement des microphones ainsi que la distance entre les micros et le type de chaque microphone. La prise de son stéréo permet de capter le champ sonore en rendant compte de la profondeur de l'image sonore. Il faut noter aussi la dépendance forte du savoir-faire des

ingénieurs du son dans le domaine de l'acoustique des salles et leur expérience d'écoute dans l'utilisation de cette technique. Même la conformité de toutes ces conditions ne garantit pas une bonne prise de son du point de vue de l'immersion sonore et de la qualité du son.

La prise de son stéréo par un couple de microphones est strictement liée à la restitution sur deux haut-parleurs. La disposition des sources acoustiques dans la scène recréée lors de la restitution, peut être trouvée à partir des lois basées sur la différence d'intensité acoustique (ou amplitude) ou de temps (ou phase). Des microphones espacés (non-coïncidents) permettent de trouver la différence de temps pour une source acoustique [6]. Il est possible de créer artificiellement une scène sonore basée sur un signal stéréo sans utiliser de microphones stéréo. Dans ce cas une technique dite « potentiomètre panoramique » (ou « pan-pot ») peut être utilisée par les ingénieurs du son pendant le mixage, et consiste à mettre en œuvre des microphones à proximité des sources sonores pour les capter de manière la plus indépendante possible. La technique « pan-pot stéréo » offre une répartition de puissance des sources sonores entre les canaux gauche et droit. Par exemple le mixage d'un concert avec cette technique permet de grouper certains instruments plutôt vers le canal gauche, d'effectuer une balance des chanteurs au centre et répartir les autres sources sonores plutôt vers la droite.

La technique classique de mixage consiste à créer une scène sonore en mixant les signaux du microphone principal et celui de chaque microphone d'appoint. Il s'agit d'ajuster le signal du microphone d'appoint dans le signal mixé, c'est-à-dire de définir les transformations d'amplitude et/ou de phase à appliquer au signal avant diffusion sur haut-parleurs, pour former une image sonore qui soit cohérente avec celle fournie par le microphone principal. La cohérence recherchée doit être spatiale, et il faut préciser pour cela la position de celui-ci en espace (azimut dans le cas « 2D », c'est-à-dire dans le plan horizontal). Elle doit être aussi temporelle, c'est-à-dire que l'on doit idéalement annuler le retard temporel entre les signaux d'appoint et les signaux du microphone principal, afin d'éviter des effets d'écho ou de coloration (filtrage en peigne). Ce retard est relié à la distance entre la source acoustique avec un microphone d'appoint associé et le microphone principal (les ondes acoustiques captées par le microphone d'appoint arrivent au microphone principal avec un retard qui est relié directement à la distance). Enfin, le dosage approprié de la source dans la scène globale se fait en ajustant le niveau du gain.

Il existe plusieurs plugins audio VST (**V**irtual **S**tudio **T**echnology de la société Steinberg – un protocole de « plugin » audio pour des logiciels audio) permettant d'effectuer un « pan-pot » de micros d'appoint avant de les mixer avec un microphone principal stéréo. Le premier représentant de cette classe de plug-in est « PanNoir » (Figure 5) qui fait partie du logiciel

Pyramix Studio de la société Merging Technologies. Avec ce plug-in, l'utilisateur est capable d'ajuster les paramètres du microphone principal (stéréo) et des microphones d'appoint en réglant manuellement le retard de chaque microphone d'appoint (correspondant à chaque source acoustique) par rapport au principal, la directivité et l'espacement des microphones (Figure 6).

Figure 5. Schéma de la captation du son et du mixage du plugin VST PanNoir (Pyramix).

Figure 6. Paramètres d'ajustement pendant le mixage par le plugin VST PanNoir.

Un autre acteur dans ce domaine est Vienna MIR Pro (Figure 7), qui propose un outil de composition de la scène sonore et de mixage. L'interface de ce plug-in permet de spatialiser directement des sources acoustiques (même en 3D, c'est-à-dire avec élévation) en ajoutant la réponse d'une salle et en modifiant l'élargissement de chaque source sonore. Ces plug-ins sont utiles dans la post-production mais ils ne sont pas capables de faire le pré-mixage en temps réel pendant un enregistrement. Dernier acteur à noter c'est Adobe Premiere (Figure 8) qui offre la possibilité de synchroniser deux pistes audio (par exemple une du microphone principal et l'autre du microphone d'appoint).

Figure 7. Plugin VST Vienna MIR PRO.

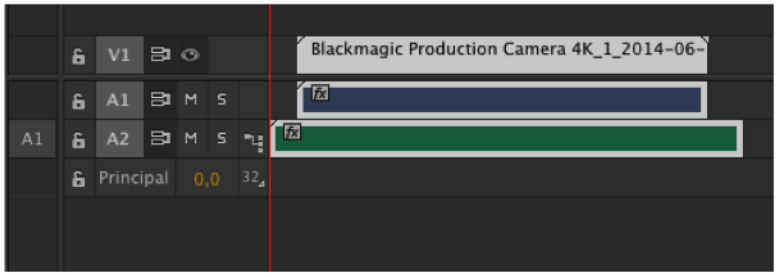

Figure 8. Interface d'Adobe Premiere. Une piste vidéo et deux pistes audio.

2.2 Prise de son multicanal

Dans les années 1970, des recherches dans le domaine de la spatialisation sonore, et des configurations d'écoute dépassant la stéréo, ont été à l'origine de la création d'un nouveau format appelé « quadriphonie » [7]. L'éditeur de musique classique Nimbus Records en Angleterre a ainsi produit les premiers disques utilisant à la fois le format stéréo et le format quadriphonique qui était décodé par la technique décrite dans les travaux de Gerzon [8]. Avec l'apparition des techniques numériques de compression audio, des nouveaux formats de stockage audio et de l'évolution des salles de cinéma, dans les années 1980 le format multicanal n'a pas tardé à émerger pour le grand public. L'évolution des systèmes de restitution vers un plus grand nombre de haut-parleurs (quadriphonie, multicanal) pour ajouter une dimension immersive, a suscité la création de nouveaux systèmes de prise de son. Un exemple de ces systèmes est « l'étoile de Williams » (Figure 9) qui se compose de 4 ou 5 microphones (et jusqu'à 7). L'évolution des systèmes de prise de son et de la diffusion du son pose un certain

nombre de difficultés techniques (augmentation du nombre de microphones pour la captation, de haut-parleurs pour la restitution, traitement du son complexe) qui sont moins perceptibles dans le système stéréo.

Figure 9. L'étoile de Williams. Arbre de microphones.

Prenons l'exemple d'un tournage de film. La prise de son s'effectue normalement sur le plateau dans un environnement sonore en même temps et au même endroit pour tous les microphones d'appoint et principal. Le microphone principal fournit une image sonore globale de la scène tout en apportant la couleur et le volume de la salle. Chaque microphone d'appoint capte en proximité une ou plusieurs sources acoustiques ciblées en les définissant plus précisément [9]. Deuxièmement, des microphones d'appoint définissent plus précisément des sources acoustiques [9]. Le placement de micro d'appoint dépend du rayonnement acoustique dans le lieu de prise de son, sauf dans les cas où l'on utilise des microphones d'appoint *idéaux* qui captent uniquement la source acoustique à proximité et ne captent pas du tout les autres sources sonores (par exemple une guitare électrique).

Les récent progrès dans les domaines des jeux vidéo, des applications multimédia et des appareils portables ouvrent de nouvelles perspectives pour la vidéo et le son. La technologie VR 360° (« Virtual Reality ») est en plein essor et la manipulation du champ sonore en trois dimension en synchronisation avec l'image est indispensable pour cette technologie. Les solutions techniques comme la « stéréophonie » ainsi que des systèmes de prise de son multicanal comme l'étoile de Williams sont moins adaptées à ces nouvelles technologies.

2.3 Prise de son ambisonique

Dans les années 1970 Gerzon présente la « périphonie » en ajoutant la dimension verticale aux dimensions horizontales de la reproduction du son autour de l'auditeur [8]. Comme système phare des approches périphoniques, Gerzon a introduit la technologie *ambisonique* qui s'adapte à différents systèmes de restitution, permet de réaliser des

manipulations de la scène sonore (rotation, focalisation, etc.) et repose sur un formalisme mathématique puissant avec une séparation claire de chaque étape du traitement de signal.

Dans la technique ambisonique, les caractéristiques spatiales du champ sonore (position angulaire des sources) sont décrites par quatre signaux : un signal correspondant à la pression acoustique mesurée à une position de référence, *O*, et trois signaux correspondant aux dérivées spatiales de la pression selon les axes *Ox*, *Oy* et *Oz*. Pour réaliser un microphone capable de capter ces signaux, Michael Gerzon a proposé [8] d'utiliser trois couples de microphone de type MS.

En effet, selon le travail de Gerzon [8] trois couples MS sont alignés suivant des axes orthogonaux avec un microphone commun « M » en remplaçant le cardioïde par un omni pour chacun des couples (Figure 10). L'espace sonore est alors capté par un micro omni (nommé la composante W) et des signaux bidirectifs (les composantes X, Y, Z). L'ensemble des quatre composantes captées par ce type de dispositif forment le format B ou autrement dit Ambisonics d'ordre 1 (Figure 10).

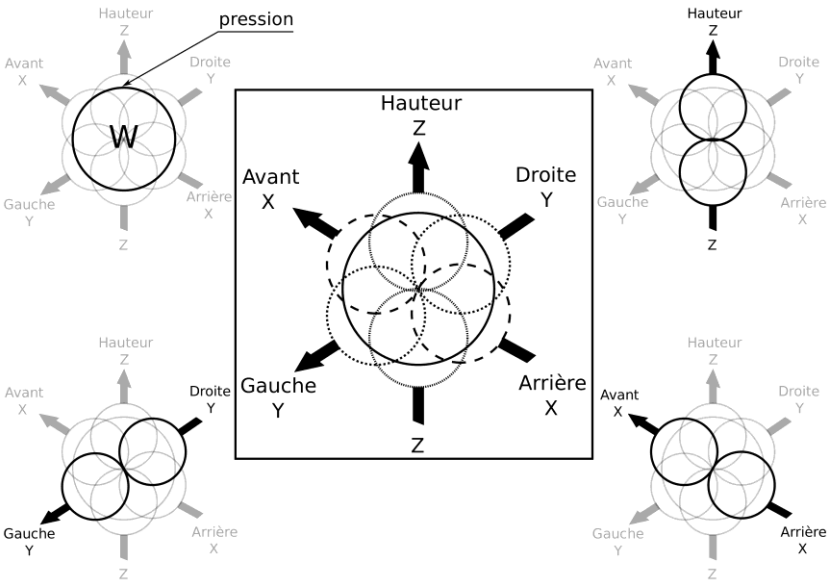

Figure 10. Format B Ambisonics d'ordre 1.

Si on considère une source acoustique émettant le signal $s(t)$ et dont la direction est décrite par le vecteur unitaire $\vec{u}$, les 4 composantes du Format B s'expriment sous la forme suivante [6]:

$$\begin{cases} w(t) = s(t) \\ x(t) = \eta \cdot s(t) \cdot \vec{u} \cdot \vec{x} \\ y(t) = \eta \cdot s(t) \cdot \vec{u} \cdot \vec{y} \\ z(t) = \eta \cdot s(t) \cdot \vec{u} \cdot \vec{z} \end{cases} \quad (1)$$

où η est un facteur de normalisation introduit par Gerzon pour conserver les rapports d'amplitudes entre les différentes composantes [8].

En pratique l'enregistrement direct des signaux au format B chacun par une capsule dédiée n'est pas très répandu (sauf dans des cas d'expérimentations) en raison des problèmes d'espacement physique des microphones et de la difficulté de les rendre coïncidents entre eux. Néanmoins, Gerzon et Craven [10] ont trouvé une solution qui permet de produire un microphone assez compact avec des capsules quasi-coïncidentes. Le format des signaux enregistrés par ce dispositif est alors nommé « Format A ». L'idée principale de ce format est de discrétiser la sphère par un tétraèdre et de placer sur chaque côté de celui-ci des microphones de type « cardioïde » (Figure 11).

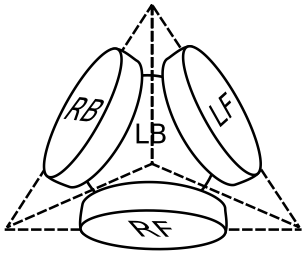

Figure 11. Format A. Placement des microphones. LF : Left-Front, RF : Right-Front, LB : Left-Back, RB : Right-Back.

Les 4 canaux LF, LB, RL, RB correspondent au format A. Le passage du format A vers le format B est réalisé par le système d'équations suivantes :

$$\begin{cases} w = LF + LB + RF + RB \\ x = LF - LB + RF - RB \\ y = LF + LB - RF - RB \\ z = LF - LB - RF + RB \end{cases} \quad (2)$$

En appliquant cette technique il a été produit toute une série de microphones sous la marque « SoundField ».

Ambisonics a ensuite évolué aux ordres supérieures donnant la naissance au HOA (pour « Higher Order Ambisonics») grâce aux travaux de J. Daniel [6], permettant d'améliorer la résolution du champ sonore (Annexe 2). Dans le cadre de cette thèse, plusieurs captations du son au format HOA ont été effectuées. Un microphone Eigenmike de la société mhAcoustics (Figure 12, [11]) a été utilisé en tant que microphone principal.

Figure 12. Microphone Eigenmike (mhAcoustics).

La carte son externe fournie par mhAcoustics a été branchée sur un ordinateur MacBook Pro avec le logiciel Reaper qui permet d'effectuer l'enregistrement jusqu'à 64 canaux simultanément. Il faut remarquer des difficultés rencontrées pendant les tests et l'utilisation du système d'enregistrement évoqué. Par exemple, l'utilisation du système sur un ordinateur PC (Windows 7) avec le logiciel Reaper n'est pas fiable et provoque souvent des artefacts audio dans l'enregistrement (perte du signal, craquement). L'utilisation de l'horloge (World Clock) fournie par la carte son mhAcoustics mène à des erreurs dans la fréquence d'échantillonnage.

Opéra « Isis et Osiris » à l'Espace de Projection (IRCAM, Paris)

L'IRCAM, en collaboration avec Radio France et b<>com, a travaillé sur l'opéra « Isis et Osiris », écrit par le compositeur Jacques Lenot et réalisé spécialement pour le système de diffusion de l'Espace de Projection (Figure 13).

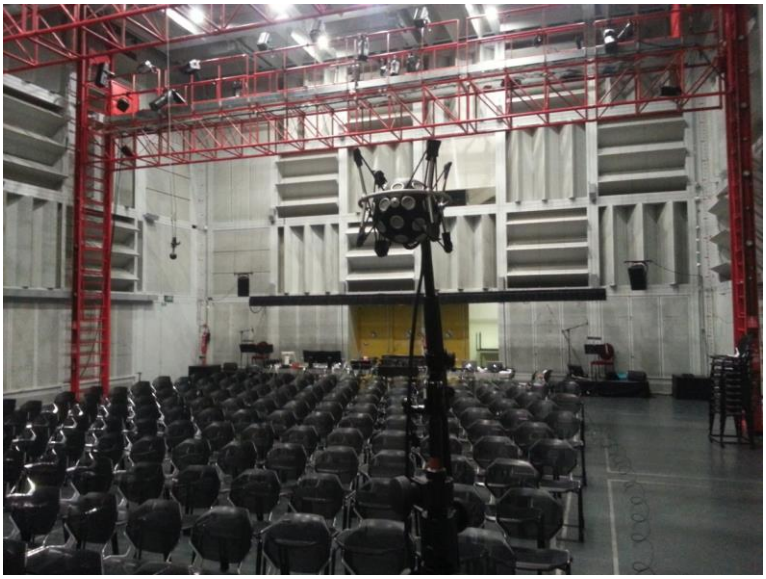

Figure 13. Espace de Projection IRCAM à Paris.

La forme et le volume de l'Espace de Projection sont modifiables grâce au plafond divisé en trois parties dont les hauteurs peuvent être réglées indépendamment. Les murs et les plafonds

sont subdivisés en modules prismatiques, dénommés périactes, présentant trois types de faces aux caractéristiques acoustiques différents (absorbantes, réfléchissantes et diffusantes). Ces périactes sont motorisés et contrôlables à distance depuis un ordinateur. La plage de temps de réverbération accessible est comprise entre 0,6 et plus de 3 secondes. L'enregistrement a été effectué avec le microphone Eigenmike et avec des microphones d'appoint à proximité de chaque musicien (Figure 14). Le microphone principal était suspendu au milieu de la salle dans la zone du public pour que l'on puisse reproduire le champ sonore et obtenir les mêmes sensations pendant la restitution que si on écoutait l'opéra avec le public.

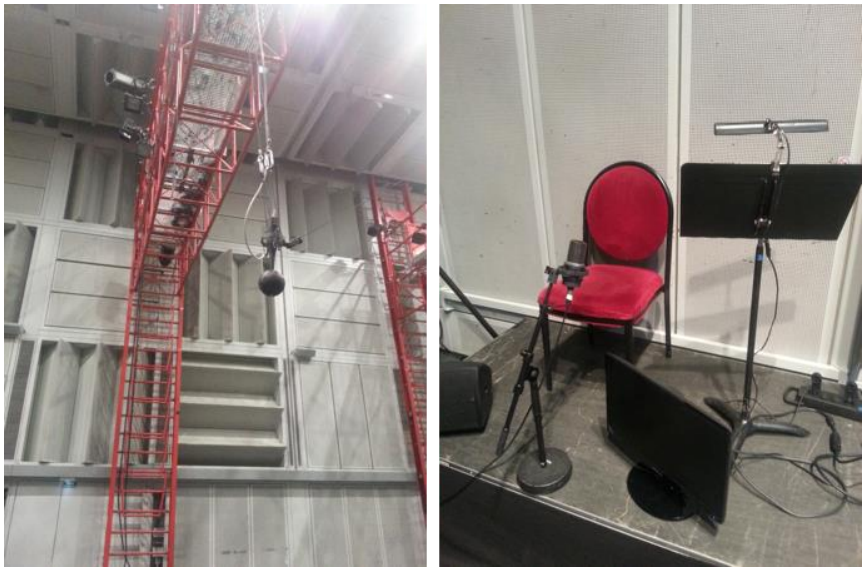

Figure 14. Espace de Projection IRCAM à Paris. Placement du microphone HOA Eigenmike (à gauche) et un des microphones d'appoint (à droite).

Le conte « La fille sans les mains » au CNSMDP

La deuxième prise de son est celle du conte, qui a été réalisée au sein du Conservatoire National Supérieur de Musique et de Danse de Paris (CNSMDP), dans le cadre d'un projet mené par des étudiants sur les prises de son immersives. Le but de ce projet est d'étudier et comparer différents systèmes de captation du son. Il avait été installé 64 microphones (Figure 15, gauche), formant des systèmes 2D, mais aussi des microphones d'appoint, ainsi que le microphone Eigenmike (Figure 15, droite). Le microphone Eigenmike était suspendu au-dessus du pupitre du chef d'orchestre. La conteuse était isolée par une structure absorbante au deuxième étage de la même salle.

Figure 15. La salle CNSMDP. Divers systèmes de captation du son.

Le projet « Tro Fanch » de Cross Channel Film Lab

Le cinéma ouvre de nouvelles perspectives pour l'utilisation de HOA en tant que microphone principal en conjonction avec des microphones d'appoint. Le projet « Tro Fanch » de Cross Channel Film Lab en collaboration avec le collectif Le Groupe Ouest et avec la participation de b<>com consiste en une série de tournages 3D (vidéo et son). Deux scènes à l'intérieur et à l'extérieur ont été réalisées par une caméra 3D et des microphones de type « mono », « stéréo » et le microphone ambisonique (Figure 16). Pour toutes les séquences, le microphone Eigenmike était en position fixe à proximité de l'action principale.

Figure 16. Tournage 3D de Tro Fanch. Scène à l'extérieur et à l'intérieur.

Court-métrage « Refuge » à Montréal, Canada

Un autre tournage a été réalisé à Montréal, Canada, en collaboration avec le studio « Jimmy Lee », et a consisté en deux scènes en extérieur dans une forêt (avec des effets sonores naturels comme la pluie et le vent) et à l'intérieur dans une église (Figure 17).

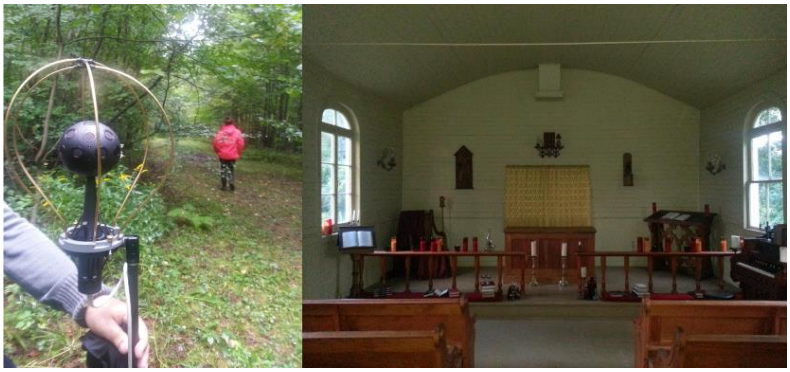

Figure 17. Tournage au Canada avec le studio Jimmy Lee.

La scène intérieure est une scène mixte entre action et musique (des cloches) mais sans microphone d'appoint. Ici le microphone Eigenmike a été utilisé comme microphone principal. Malgré des difficultés techniques, un système d'enregistrement mobile (un ordinateur MacBook Pro, deux batteries « Tekkeon myPower ALL » pour alimenter la carte son mhAcoustics et la boîte de synchronisation « Grass Valley ADVC G4 ») a été créé pour satisfaire aux conditions d'usage complexes (Figure 18).

Figure 18. Système de prise de son HOA pendant le tournage au Canada.

2.4 Assistance automatique au mixage HOA

La problématique d'assistance automatique au mixage du son immersif au format HOA (dans la suite du document « mixage HOA »), adressée dans ce travail de thèse, peut être décrite comme une partie de la chaîne de production de contenus « audio immersifs » (Figure 3). L'estimation des paramètres de mixage effectuée manuellement par l'ingénieur du son est

fastidieuse et souvent approximative. La solution proposée ici facilite le travail de l'ingénieur du son en proposant une estimation automatique des paramètres souhaités.

Notre approche automatisée comprend un module (bloc d'analyse sur la Figure 19) d'estimation des paramètres principaux (retard, azimut, élévation et gain) pour fournir des paramètres estimés à l'ingénieur du son qui peut valider, corriger ou ajuster l'estimation afin d'effectuer le mixage. A l'entrée du système proposé se trouvent les signaux de chaque microphone d'appoint (mono ou stéréo, nommés A1 à An dans la Figure 19) et le signal du microphone principal au format HOA.

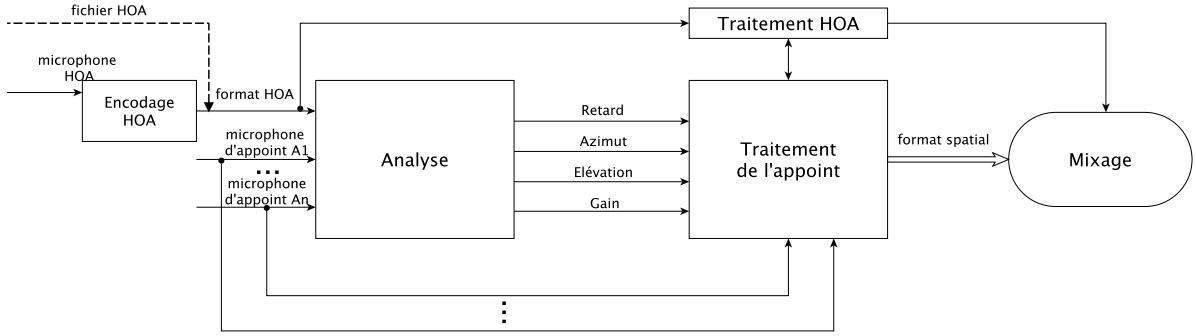

Figure 19. Estimation de paramètres et mixage au format spatial HOA.

Le problème à résoudre est d'effectuer un mixage « sans artefact » entre des signaux d'une représentation audio spatiale multicanale correspondant à un micro principal, et ceux de micros d'appoint. Ce mixage comprend une étape d'encodage spatial de chaque signal d'appoint vers une représentation du même type que celle associée au micro principal. Pour éviter des artefacts, il est nécessaire :

- d'appliquer un retard temporel (éventuellement négatif dans le cas où par exemple ils n'ont pas été enregistrés sur la même station) sur chaque micro d'appoint afin de réduire le décalage temporel entre un signal caractéristique présent à la fois dans le micro d'appoint et le micro principal.
- que les paramètres d'encodage spatial (paramètres de localisation, position) du signal d'appoint soient conformes avec l'image de la scène captée par le micro principal

Il existe des outils qui gèrent le format HOA en proposant un ensemble de traitements pour le premier ordre et les ordres supérieurs : encodage spatial, manipulation de la scène sonore et décodage mais aucun de ces outils ne proposent de fonctions de pré-mixage automatique. Les outils identifiés sont :

- les plugins VST d'Orange Labs (non commercialisés). Le plugin HOAmicProcessor propose un encodage vers HOA des signaux issus du micro Eigenmike. Le plugin HOARotator permet d'appliquer des rotations à une scène HOA, le plugin

HOAEncoder permet d'encoder en HOA une source mono, le plugin HOASpkDecoder permet de décoder le HOA pour un système de haut-parleurs, et le plugin HOABinDecoder permet de décoder pour le casque en binaural. Ils sont proposés avec une interface graphique.

- les plugins Ambix du développeur Matthias Kronlachner [12]. Cette suite de plugins propose un grand nombre de traitements (binaural, decoder, format converter, directional loudness, encoder, maxre, mirror, rotation, vmic, warp, widening). Ces plugins nécessitent des « binaural decoder presets » qui incluent des matrices de décodage et des BRIR (« Binaural Room Impulse Responses »). Il n'y a pas de « preset generator » d'inclus (matrice de decodage). Il est conseillé d'utiliser la Ambisonic Decoder Toolbox d'Aaron Heller.
- les plugins TOA (Third Order Ambisonic) de « Blue Ripple Sound » [13] gérant jusqu'au 3ème ordre HOA. L'ensemble « Core » fournit une librairie gratuite de plugins de base (panners, decoders, beamer, virtual microphone, rotation, metering, visualisers), et d'autres plugins payants plus spécifiques sont également proposés (decoders, Harpex upsampler, manipulateurs, reverb, upmixers).
- Le CICM (Centre de recherche Informatique et Création Musicale de l'Université Paris 8) propose une bibliothèque HOA [14] qui comprend le rendu 3D. La bibliothèque HOA est gratuite, open-source et mise à disposition par le CICM. Il s'agit d'une suite de plugins de décodage pour 5, 6, 8, 16 haut-parleurs et pour le casque (binaural). Ces plugins sont à appliquer sur des signaux mono. Ils appliquent à la fois l'encodage et le décodage HOA. Ils proposent une interface graphique avec vue de dessus pour placer les sources.
- les plugins Spatium [15] constituent une suite gratuite, open-source et modulaire d'outils pour la spatialisation.
- Les outils gratuits de spatialisation du son pour la réalité virtuelle Two Big Ears' "Spatial Workstation" [16].

2.5 Conclusion

Dans ce chapitre nous avons décrit les problématiques liées aux techniques de prise de son stéréo et HOA. L'état de l'art retrace l'historique des domaines du mixage et de la prise de son classique et HOA sur laquelle se base la thèse. Enfin l'immersion sonore a été définie d'un point de vue perceptif.

Le sujet du mixage du son classique et HOA montre la problématique globale de l'identification des paramètres de mixage. Il explique les difficultés associées à chaque étape du mixage classique.

Plusieurs prises de son au format HOA avec des microphones d'appoint ont été réalisées à l'aide d'un microphone HOA. A partir de ces contenus et méthodes proposées dans le chapitre 3, un algorithme du mixage HOA avec des signaux des microphones d'appoint a été développé (chapitre 4) à la base duquel il a été effectué plusieurs tests de performance (chapitre 5).

Le premier ordre Ambisonics a été introduit sur lequel se base les méthodes développées (chapitre 3). Des outils de mixage du son au format stéréo ont été illustrés dans ce chapitre ainsi que le manque des outils de mixage au format HOA.

Le but du travail de cette thèse « l'assistance automatique au mixage HOA » a été évoqué ainsi que les étapes nécessaires afin d'effectuer le mixage HOA.

3 Estimation des paramètres de sources acoustiques

3.1 Problème posé et état de l'art technique

L'assistance automatique au mixage vise à proposer les paramètres permettant de mixer les signaux captés par le microphone principal avec ceux des microphones d'appoint, sans introduire d'artefact audible. L'extraction de ces paramètres se base sur une représentation géométrique des sources acoustiques et des microphones. Vue du microphone principal (considéré comme ponctuel), une source acoustique, considérée comme un point dans l'espace 3D, est déterminée par les paramètres suivants (Figure 20):

- distance (r) entre le microphone principal et la source
- position décrite par l'azimut (θ) et l'élévation (φ) de la source

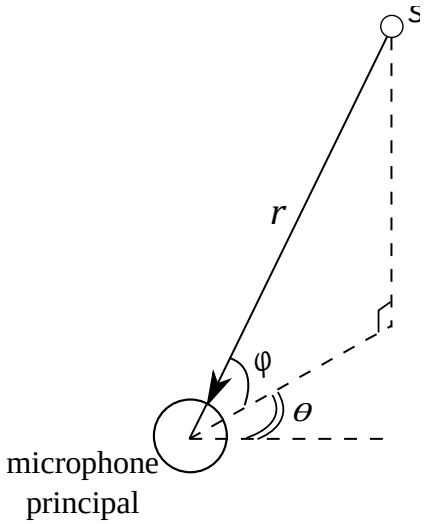

Figure 20. Source acoustique en espace 3D avec un microphone principal. « s » - source acoustique. « r, θ, φ » - paramètres décrivants la position de la source en espace 3D.

Le mixage d'une source acoustique peut être effectué à partir de l'azimut, de l'élévation et du retard τ entre le signal émis par la source acoustique et le signal capté par le microphone principal. En termes acoustiques cette valeur τ approche le temps de propagation des ondes acoustiques provenant de la source (point « s » sur la Figure 20) jusqu'au capteur acoustique (« microphone principal » sur la figure). Donc le retard τ peut être facilement exprimé à partir de la distance r (Figure 20): $\tau = r/c$, où c est la vitesse du son (≈340 m/s dans l'air).

Dans un système sans bruit et sans réverbération, comme celui illustré sur la Figure 20, le signal capté par le microphone principal (dénoté « p_{pr} » comme la « **p**ression » avec l'indice « pr » relié au microphone **pr**incipal) correspond à la version retardée de τ du signal « source ».

Pour exprimer le signal capté par le microphone principal et retardé par rapport au signal de la source acoustique on utilise l'opération classique de convolution du signal de la source acoustique par un filtre de réponse impulsionnelle h :

$$p_{\mathrm{pr}}(t) = [h * s](t) \tag{3}$$

où le filtre h correspond à une impulsion à l'instant τ avec un certain niveau du gain g:

$$h = g \cdot \delta(t - \tau) \tag{4}$$

En pratique on ne dispose pas du signal de la source acoustique mais du signal capté par un microphone d'appoint à proximité de cette source. Dans ce cas (Figure 21) si la distance entre le microphone d'appoint a et la source acoustique s est négligeable on peut considérer le signal de microphone d'appoint comme le signal direct de la source acoustique et, par conséquent, le signal du microphone principal s'exprime avec les mêmes équations (3) et (4) en remplaçant $s(t)$ par $a(t)$.

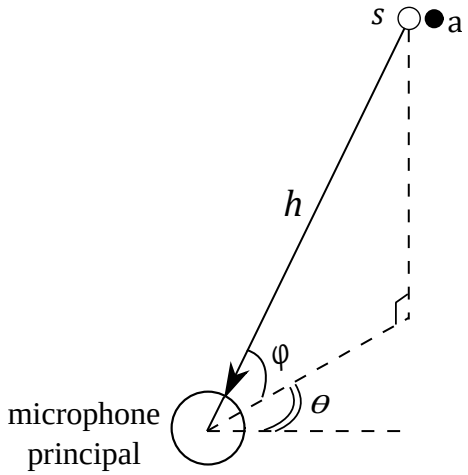

Figure 21. Source acoustique en espace 3D avec un microphone principal. « *s* » - source acoustique, « *a* » - microphone d'appoint.

Les deux exemples jusqu'à présent correspondent à la situation idéale dans laquelle seule une source est présente, la propagation acoustique se fait en champ libre (absence de réverbération, le signal capté par le microphone principal correspond au « trajet direct »), les signaux microphoniques sont dénués de bruit et la distance entre la source et le microphone d'appoint est négligeable. La plupart du temps, les conditions dans lesquelles la prise de son s'effectue ne sont pas idéales et il faut prendre en compte les paramètres décrivant la scène sonore complète. Cette description de la scène sonore complète contient, tout d'abord, des informations sur les ondes acoustiques secondaires (réverbération acoustique, dit « trajet indirect »), mais aussi sur la présence éventuelle de bruit intrinsèque (bruit de mesure des microphones) et la distance significative entre la source acoustique et le microphone d'appoint.

Dans le cas d'une scène sonore complète, on peut décomposer le filtre modélisant la propagation acoustique entre la source *m* et le micro *Mic* (qui peut être, soit le microphone principal « pr », soit un des microphones d'appoint « a_n »), dénoté $h_{s_m,Mic}$, en deux parties correspondant aux trajets direct et indirect des ondes acoustiques, respectivement :

$$h_{s_m,Mic} = h_{s_m,Mic}^{(direct)} + h_{s_m,Mic}^{(indirect)} \quad (5)$$

Par analogie avec l'équation (4) dans le cas avec des ondes acoustiques directes et indépendantes de la fréquence, la réponse impulsionnelle décrivant la propagation directe, $h_{s_m,Mic}^{(direct)}$, est définie par l'expression suivante:

$$h_{s_m,Mic}^{(direct)} = g_{s_m,Mic} \cdot \delta(t - \tau_{s_m,Mic}) \quad (6)$$

où $\tau_{s_m,Mic}$ et $g_{s_m,Mic}$ sont, respectivement, le retard de la propagation et le gain correspondant à la propagation acoustique entre la source m et le microphone Mic.

On peut en déduire, par analogie avec le cas d'une source et un microphone (3), le signal d'un microphone (soit principal, soit d'appoint) qui capte m sources acoustiques dans la scène avec n microphones d'appoint et un microphone principal (Figure 22) :

$$p_{Mic}(t) = \sum_{m=1}^{M} \left\{ \left[h_{s_m,Mic}^{(direct)} * s_m \right](t) + \left[h_{s_m,Mic}^{(indirect)} * s_m \right](t) \right\} + v_{Mic}(t) \quad (7)$$

où « * » est l'opération de convolution et $v_{Mic}(t)$ est le signal correspondant au bruit intrinsèque du microphone.

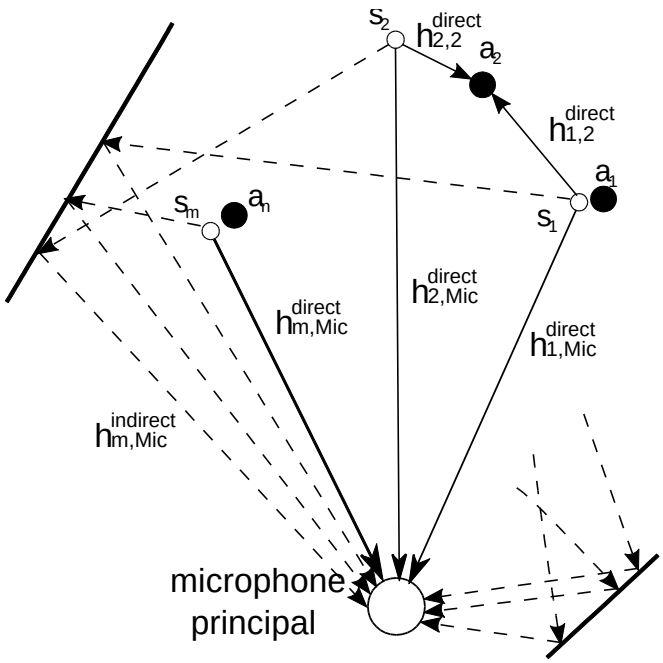

Figure 22. Scène sonore. m sources acoustiques « s » et n microphones d'appoint « a ». Les lignes en trait plein sont associées aux trajets directs de source acoustique, les lignes à tirets longs montrent les trajets indirects (réverbération).

Ainsi, si on considère une scène sonore avec une source acoustique et un microphone d'appoint associé, le signal $p_{a_1}(t)$ capté par le microphone d'appoint « a_1 » s'écrit sous la forme :

$$p_{a_1}(t) = \left[g_{s_1,a_1} \cdot \delta\left(t - \tau_{s_1,a_1}\right) * s_1\right](t) = g_{s_1,a_1} \cdot s_1\left(t - \tau_{s_1,a_1}\right) \qquad (8)$$

Les indices « s_1, a_1 » dans l'équation (8) montrent le lien entre la source et le microphone. En particulier, τ_{s_1,a_1} est un retard entre la source acoustique « s_1 » et le signal capté par le microphone d'appoint « a_1 ». Par analogie avec (8) on peut exprimer le signal capté par le microphone principal « pr » :

$$p_{\mathrm{pr}}(t) = \left[g_{s_1,\mathrm{pr}} \cdot \delta\left(t - \tau_{s_1,\mathrm{pr}}\right) * s_1\right](t) = g_{s_1,\mathrm{pr}} \cdot s_1\left(t - \tau_{s_1,\mathrm{pr}}\right) \qquad (9)$$

En pratique, le signal de la source acoustique même ($s_1(t)$ dans l'équation (8)) n'est pas connu, seul le signal de la source capté par le microphone d'appoint est disponible. On peut exprimer la source à partir du signal du micro d'appoint à partir de l'équation (8):

$$s_1(t) = \frac{1}{g_{s_1,a_1}} p_{a_1}\left(t + \tau_{s_1,a_1}\right) \qquad (10)$$

A l'aide des équations (9) et (8), il est possible d'exprimer le signal capté par le microphone principal $p_{\mathrm{pr}}(t)$ à partir du signal capté par le microphone d'appoint $p_{a_1}(t)$:

$$p_{\mathrm{pr}}(t) = \frac{g_{s_1,\mathrm{pr}}}{g_{s_1,a_1}}\left[\delta\left(t - \tau_{s_1,\mathrm{pr}} + \tau_{s_1,a_1}\right) * p_{a_1}\right](t) \tag{11}$$

En pratique, il est difficile de trouver les retards $\tau_{s_m,\mathrm{pr}}$ et τ_{s_m,a_m} en utilisant seulement les signaux captés par les microphones. Le signal de la source acoustique s_m est capté par le microphone d'appoint a_n avec le même retard τ_{s_m,a_m} en tous les points d'un cercle (Figure 23, lignes à tirets) centré sur la source s_m et de rayon égal à la distance entre la source acoustique et le microphone d'appoint. La source acoustique située dans la direction de la source acoustique s_m pour le microphone principal peut être considérée comme une « source apparente » (Figure 23, en pointillés). La source apparente est localisée à l'intersection du cercle correspondant au retard source-appoint et de la droite reliant la source au microphone principal. Le retard entre le signal du microphone principal et celui du microphone d'appoint est donc la différence entre le retard « source-appoint » et le retard « source-principal » (H_n^{direct} sur la Figure 23).

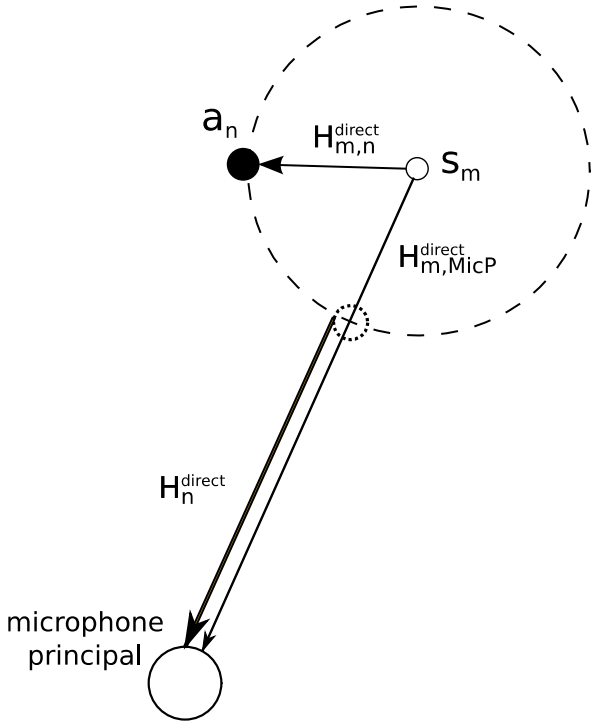

Figure 23. Source apparente. Propagation des ondes acoustiques avec une source acoustique, un microphone d'appoint et le microphone principal.

Ce retard est proportionnel à la distance entre les microphones uniquement si la distance entre la source et le microphone d'appoint est négligeable et, dans le cas général, le retard s'exprime :

$$\tau_n = \tau_{s_m,\mathrm{pr}} - \tau_{s_m,a_n} \tag{12}$$

3.1.1 Retard

Le retard entre deux signaux peut être estimé en calculant l'intercorrélation entre ces signaux. L'observation du maximum d'intercorrélation révèle le décalage entre les deux signaux. L'intercorrélation entre deux signaux $x(t)$ et $y(t)$ *en* temps discret est donnée par :

$$ {}_{\tau_1}\langle x|y\rangle_{\tau_2} = \sum_{k=K_1}^{K_2} x(k-\tau_1)\cdot y(k-\tau_2) \tag{13} $$

où les signaux x et y sont retardés par τ_1 et τ_2 ; K_1 et K_2 sont les extrémités de la fenêtre temporelle considérée.

On peut déduire les propriétés suivantes de l'équation (13) :

- $\langle x|y\rangle_\tau = {}_0\langle x|y\rangle_\tau$
- $\|x\|_\tau = \sqrt{{}_\tau\langle x|x\rangle_\tau}$ (norme de x sur la fenêtre temporelle considérée)

Dans la suite du document le retard entre deux signaux est estimé à l'aide de la fonction d'intercorrélation normalisée, qui s'exprime comme suit :

$$ C(\tau) = \frac{{}_\tau\langle x|y\rangle}{\|x\|_\tau \cdot \|y\|} \tag{14} $$

Le retard estimé est la valeur de retard pour laquelle la fonction $C(\tau)$ atteint sa valeur maximale :

$$ \tilde{\tau} = \underset{\tau}{\operatorname{argmax}}\, C(\tau) \tag{15} $$

Une estimation précise du retard permet de mixer le signal enregistré par le microphone d'appoint avec celui provenant du microphone principal afin de préciser la source acoustique dans le mix final. L'estimation par la fonction (14) ne donne pas de bons résultats dans un environnement bruité ou réverbérant. A titre d'exemple la corrélation d'un signal périodique avec sa version décalée présente de nombreux pics (voir Figure 24), ce qui rend l'estimation du retard difficile. La présence de sources multiples peut également poser problème. Considérons le cas où la scène comporte deux sources acoustiques (Figure 25).

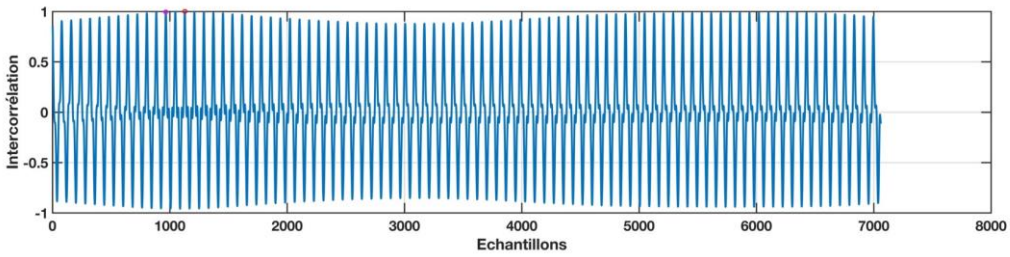

Figure 24. Corrélation entre le signal périodique capté par le microphone d'appoint et le signal capté par le microphone principal dans une scène sonore avec une source acoustique. Deux points rouges montrent les deux pics maximaux.

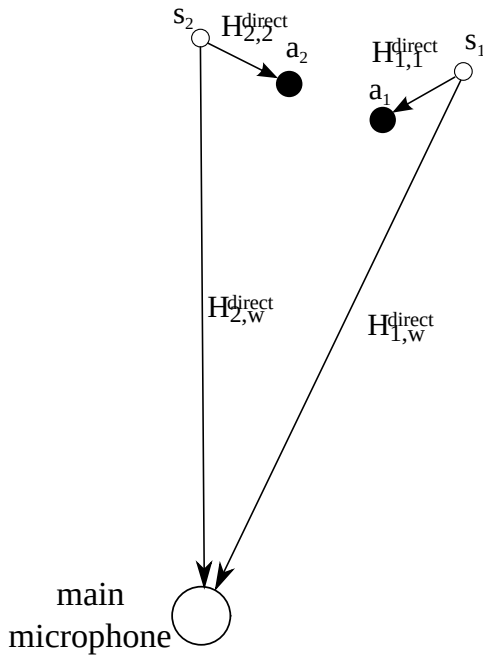

Figure 25. Deux sources acoustiques avec un microphone principal.

Le signal de chaque micro d'appoint peut être exprimé en fonction du signal de la source acoustique correspondante :

$$a_m(t - \tau_{m,m}) = g_{m,m} \cdot s_m(t) \qquad (16)$$

où $s_m(t)$ est le signal émis par la source m et $g_{m,m}$ est le gain associé correspondant au micro d'appoint a_m. En faisant l'hypothèse que chaque microphone d'appoint ne capte que la source à laquelle il est associé, les signaux $s_1(t)$ et $s_2(t)$ peuvent donc être exprimés comme suit :

$$s_1(t) = {}^{1}/_{g_{1,1}} \cdot a_1(t + \tau_{1,1})$$

$$s_2(t) = {}^{1}/_{g_{2,2}} \cdot a_2(t + \tau_{2,2}) \tag{17}$$

Si on considère le microphone principal comme le microphone HOA, la première composante HOA du champ sonore s'écrit :

$$w(t) = g_{1,w} \cdot s_1(t - \tau_{1,w}) + g_{2,w} \cdot s_2(t - \tau_{2,w}) \tag{18}$$

En utilisant l'équation (17) on peut exprimer $s_1(t)$ et $s_2(t)$ en fonction de $a_1(t)$ et $a_2(t)$. La composante $w(t)$ peut alors être écrite sous la forme suivante :

$$w(t) = {}^{g_{1,w}}/_{g_{1,1}} \cdot a_1(t - \tau_{1,w} + \tau_{1,1}) + {}^{g_{2,w}}/_{g_{2,2}} \cdot a_2(t - \tau_{2,w} + \tau_{2,2}) \tag{19}$$

Avec les définitions suivantes :

$$\tau_1 = \tau_{1,w} - \tau_{1,1}; \quad \tau_2 = \tau_{2,w} - \tau_{2,2}; \quad g_1 = {}^{g_{1,w}}/_{g_{1,1}}; \quad g_2 = {}^{g_{2,w}}/_{g_{2,2}}; \tag{20}$$

la composante $w(t)$ s'exprime :

$$w(t) = g_1 \cdot a_1(t - \tau_1) + g_2 \cdot a_2(t - \tau_2) \tag{21}$$

Le produit scalaire entre la première composante HOA, $w(t)$, et le signal du microphone d'appoint $a(t)$ décalé de τ peut être déduit de l'équation (13):

$$\begin{aligned} \langle w(t)|a_1(t)\rangle_\tau &= \langle a_1(t)|w(t)\rangle_{-\tau} \\ &= \langle a_1(t)|g_1 \cdot a_1(t - \tau_1 + \tau)\rangle + \langle a_1(t)|g_2 \cdot a_2(t - \tau_2 + \tau)\rangle \end{aligned} \tag{22}$$

$$\begin{aligned} \langle w(t)|a_2(t)\rangle_\tau &= \langle a_2(t)|w(t)\rangle_{-\tau} \\ &= \langle a_2(t)|g_1 \cdot a_1(t - \tau_1 + \tau)\rangle + \langle a_2(t)|g_2 \cdot a_2(t - \tau_2 + \tau)\rangle \end{aligned} \tag{23}$$

où τ est une variable temporelle.

Deux cas peuvent être distingués, selon que les deux sources sont décorrélées ou corrélées. Dans le premier cas, le terme $\langle a_1(t)|g_2 \cdot a_2(t - \tau_2 + \tau)\rangle$ dans l'équation (22) peut être négligé par rapport à $\langle a_1(t)|g_1 \cdot a_1(t - \tau_1 + \tau)\rangle$ à condition que le fenêtre temporelle soit suffisamment longue. Idem pour l'équation (23). Dans le cas où les sources sont décorrélées les équations (22) et (23) deviennent donc :

$$\langle a_1(t)|w(t)\rangle_{-\tau} \approx \langle a_1(t)|g_1 \cdot a_1(t-\tau_1+\tau)\rangle \quad (24)$$

$$\langle a_2(t)|w(t)\rangle_{-\tau} \approx \langle a_2(t)|g_2 \cdot a_2(t-\tau_2+\tau)\rangle \quad (25)$$

Pour la valeur $\tau = \tau_1$ (resp. τ_2) on obtient la corrélation maximale pour l'équation (24) (resp. (25)) puisque la fonction d'autocorrélation est maximale en 0.

Dans le cas où les sources sont corrélées, la fonction d'intercorrélation peut présenter deux pics correspondants aux 2 retards τ_1 et τ_2 si les signaux sont transitoires (Figure 26 (haut et milieu)) et des pics périodiques si les signaux le sont (Figure 26 (bas)). Dans ce cas, la détection du décalage entre les signaux $w(t)$ et $a_1(t)$ en cherchant la position du pic maximal devient peu robuste.

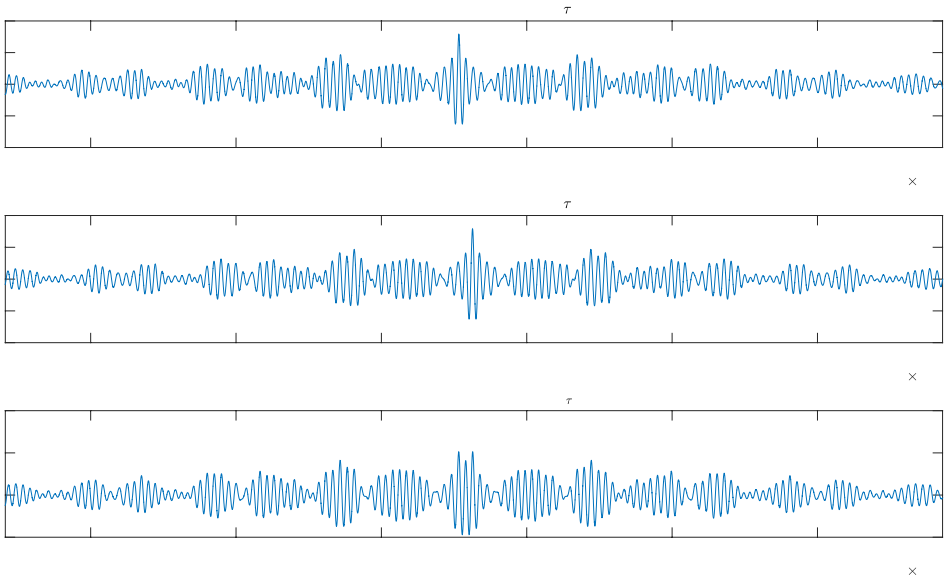

Figure 26. Sources corrélées. Cross-corrélation entre : (haut) entre $a_1(t)$ et $a_1(t)$, (milieu) entre $a_1(t)$ et $a_2(t)$, (bas) corrélation finale$\langle w(t)|a_1(t)\rangle_\tau$, simplement la sommation de deux premières qui montre les deux pics maximaux de la même amplitude.

Jusqu'à présent on a considéré le cas d'une scène sonore constituée de deux sources acoustiques, en l'absence de réverbération et sans effet de diaphonie (chaque microphone d'appoint ne capte que la source sonore à laquelle il est associé). On a montré que l'estimation du retard par la recherche du maximum d'intercorrélation est efficace lorsque les sources acoustiques sont décorrélées. En revanche, dans le cas de deux sources acoustiques corrélées l'estimation est moins robuste.

Il est nécessaire de concevoir une méthode d'estimation du retard plus évoluée afin de réduire les erreurs d'estimation dans le cas d'une scène sonore plus complexe. Dans le domaine de la localisation de sources acoustiques il existe un certain nombre de méthodes classiques

basées sur le traitement du son capté par deux microphones au minimum (Figure 27). La famille de méthodes d'estimation la plus commune est celle des méthodes dites de TDOA (« Time difference of arrival »), Le principe de ces méthodes se fonde sur la mesure du premier front d'onde acoustique par deux capteurs[17].

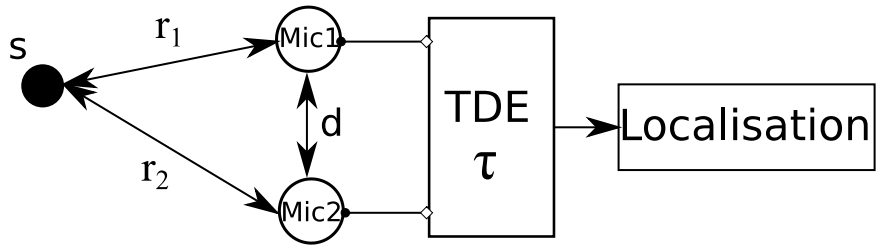

Figure 27. « Time Difference Estimation » (TDE) algorithme. « S » - source acoustique; « Mic1 », « Mic2 » - microphones.

A partir du premier front d'onde il est possible d'estimer la différence de temps de propagation (TDOA) et la différence de gain entre les deux microphones. Pour le système présenté sur la Figure 27 la différence de temps de propagation est proportionnelle à la différence entre les distances de chaque microphone à la source acoustique [18]:

$$\Delta\tau \sim (r_1 - r_2) \tag{26}$$

où « ~ » indique une relation de la proportionnalité.

Par ailleurs, le ratio des puissances des signaux microphoniques est proportionnel au ratio entre les distances de chaque microphone vers la source acoustique [18] au carré :

$$\mu \sim \frac{r_2^2}{r_1^2} \tag{27}$$

Dans le cas général [19] une paire de microphones p est constituée des microphones $\{l, k\}$ où $l, k \in [1, \dots, M], k \neq l$. Avec la position de la paire p (m_l et m_k) et la position de la source acoustique r_i la TDOA $\Delta\tau_{p,r_i}$ entre la paire s'exprime :

$$\Delta\tau_{p,r_i} = (D(r_i, m_l) - D(r_i, m_k)) \cdot c^{-1} \tag{28}$$

où $D(\bullet, \bullet)$ est la distance entre la source et le microphone.

Déduire la position de la source de la valeur de TDOA constitue un problème dit « inverse ». Dans ce cas précis, le problème est mal posé puisqu'il existe une infinité de positions de sources pouvant expliquer la valeur de TDOA observée (surface d'une hyperboloïde). Une manière possible de supprimer cette ambiguïté est d'utiliser plusieurs paires de microphones [20]. La position de la source (ou des sources) peut alors être estimée par la méthode des moindres carrés ou bien encore en sommant les hyperboles correspondant aux données provenant des différents microphones [21].

3.1.2 Position et gain

Dans le domaine de localisation des sources acoustiques il faut noter quelques résultats. Par exemple, dans le domaine stéréo l'auteur de l'article [22] décrit une technique d'extraction de paramètres de mixage (azimut et gain) à partir des différentes pistes audio d'entrée ainsi que des pistes du mixage final [21]. Un schéma décrivant le principe de base du mixage stéréo est présenté sur la Figure 28. Les pistes audio d'entrée, x_k, sont amplifiées par les gains g_k et retardées par τ_k. Un panning d'amplitude, exprimé par les gains p_{Lk}, p_{Rk} est ensuite appliqué aux signaux d'entrée. Le but de la technique présentée dans [22] est de déterminer les gains devant être appliqués à chaque piste audio d'entrée pour reproduire les canaux gauche et droite (L/R) du mixage stéréo final.

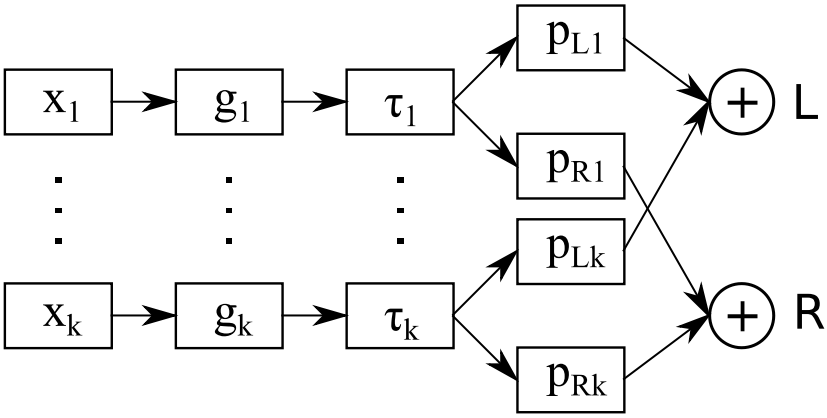

Figure 28. Mixage stéréo avec les paramètres estimés.

Les gains recherchés sont le produit des gains appliqués avant et pendant le mixage des pistes d'entrée. Ils sont donc donnés par les relations suivantes :

$$\alpha_L = g \cdot p_L$$
$$\alpha_R = g \cdot p_R \quad (29)$$

Or, si on utilise la loi de panning VBAP (« Vector Base Amplitude Panning », voir Annexe 1), la somme des carrés des gains de panning est égale à 1 et on a :

$$p_L^2 + p_R^2 = 1 \quad (30)$$

$$p_L = \cos(\theta), \quad p_R = \sin(\theta) \quad (31)$$

où θ est l'angle correspondant au mixage stéréo, compris dans l'intervalle $[0, \pi/2]$.

Pour une piste audio donnée on peut donc déduire le gain g des équations (29) et (30):

$$\sqrt{\alpha_L^2 + \alpha_R^2} = \sqrt{g^2 p_L^2 + g^2 p_R^2} = \sqrt{g^2 (p_L^2 + p_R^2)} = g \quad (32)$$

Une fois le gain g calculé on peut calculer les valeurs des gains de panning grâce à l'équation (29). Finalement on déduit la valeur de l'angle θ de l'équation (31) :

$$\theta = \arccos(p_L) = \arcsin(p_R), \quad \theta \in [0, \pi/2] \tag{33}$$

Cette technique est relativement proche de celle proposée dans cette thèse (chapitre 3.2.2) mais elle n'est pas applicable dans le domaine HOA et pour les cas plus complexes avec le bruit ou la réverbération. Par ailleurs elle est assez sensible à la présence de bruit et de réverbération ce qui la rend inutilisable dans le cas d'une scène complexe.

Une autre technique mettant en œuvre la localisation de sources acoustiques est la méthode DirAC (Directional Audio Coding) proposée par Pulkki [23]. L'étape d'analyse des signaux employée dans la méthode DirAC est illustrée sur la Figure 29. Les signaux Ambisonic d'ordre 1 sont tout d'abord séparés en bandes de fréquences, par exemple à l'aide d'une transformée de Fourier à court terme (STFT, « Short Time Fourier Transform ») ou d'un banc de filtres QMF (filtre miroir en quadrature, plus détaillé en[23]). Pour chaque bande de fréquences les signaux sont ensuite analysés de manière à estimer une position de source acoustique.

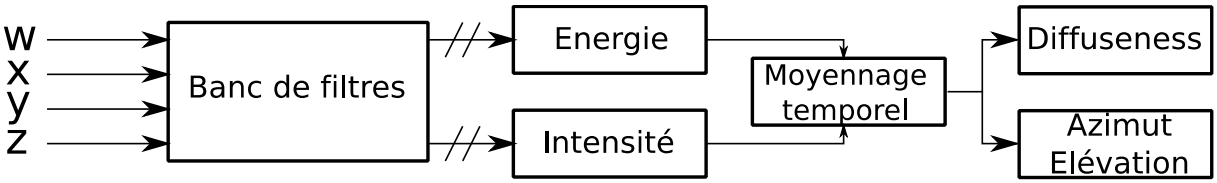

Figure 29. Analyse DirAC. B-format à l'entrée. En sortie du système les paramètres estimés : azimut, élévation et diffusivité.

L'estimation de la position de la source acoustique repose sur la notion d'intensité acoustique, qui représente le flux d'énergie sonore en un point du champ acoustique. L'intensité acoustique mesurée au point $\boldsymbol{r}$ et à l'instant t est donnée par :

$$I(\boldsymbol{r}, t) = p(\boldsymbol{r}, t) \cdot \boldsymbol{v}(\boldsymbol{r}, t) \tag{34}$$

Dans le domaine fréquentiel l'intensité acoustique est complexe et s'exprime :

$$\boldsymbol{\Pi} = \frac{1}{2} p \boldsymbol{v}^* = \boldsymbol{I} + j\boldsymbol{J} \tag{35}$$

où le vecteur $\boldsymbol{I}$ est *l'intensité active*, correspondant à une propagation de l'énergie d'un point de l'espace à un autre, et $\boldsymbol{J}$ est *l'intensité réactive* qui correspond aux fluctuations locales n'impliquant pas de propagation (ondes stationnaires).

Une autre notion intervenant dans l'estimation de la position de la source est celle de *vecteur vélocité* **V**, qui est donnée par le rapport de la vitesse particulaire sur la pression acoustique :

$$\mathbf{V}(\boldsymbol{r}) = -\frac{1}{\rho c} \frac{\boldsymbol{v}(\boldsymbol{r}, t)}{p(\boldsymbol{r}, t)} \tag{36}$$

Le vecteur vélocité **V** est complexe et peut être représenté avec ses parties réelle et imaginaire :

$$\begin{cases} \boldsymbol{\Omega}(\boldsymbol{r}) = \Re\big(\boldsymbol{V}(\boldsymbol{r})\big) = r_V \boldsymbol{u}_V \\ \boldsymbol{\Phi}(\boldsymbol{r}) = \Im\big(\boldsymbol{V}(\boldsymbol{r})\big) = r_\Phi \boldsymbol{u}_\Phi \end{cases} \quad (37)$$

Comme pour l'intensité acoustique, la partie réelle $\boldsymbol{\Omega}(\boldsymbol{r})$ a trait à la propagation sonore tandis que la partie imaginaire $\boldsymbol{\Phi}(\boldsymbol{r})$ correspond aux fluctuations locales du champ acoustique [6].

La vitesse particulaire peut être déduite des composantes Ambisonics d'ordre 1[24] :

$$\boldsymbol{v}(t) = -\frac{1}{Z_0\sqrt{2}}(x(t)\boldsymbol{e}_x + y(t)\boldsymbol{e}_x + z(t)\boldsymbol{e}_x) \quad (38)$$

où Z_0 est une impédance acoustique qui est égale au produit de la masse volumique moyenne de l'air ρ_0 par la vitesse du son c.

Par ailleurs la pression acoustique est donnée par le signal $w(t)$, on a donc :

$$p(t) = w(t) \quad (39)$$

En substituant p et $\boldsymbol{v}$ par les expressions des équations (38) et (39) l'intensité acoustique (34) peut être exprimée en fonction des signaux Ambisonic:

$$\boldsymbol{I}(t) = -\frac{w(t)}{Z_0\sqrt{2}}(x(t)\boldsymbol{e}_x + y(t)\boldsymbol{e}_x + z(t)\boldsymbol{e}_x) \quad (40)$$

Dans la méthode DirAC la direction de la source correspond à la direction de provenance du flux d'énergie acoustique, il s'agit donc de la direction indiquée par l'opposé du vecteur $\boldsymbol{I}(t)$. Le facteur $\frac{1}{Z_0\sqrt{2}}$ est négligé car il est sans relation avec la direction. L'azimut et l'élévation peuvent donc être exprimés en fonction des signaux Ambisonic comme suit[23]:

$$\tilde{\theta} = \begin{cases} \arctan\left(\dfrac{\langle y(t)|w(t)\rangle}{\langle x(t)|w(t)\rangle}\right), & \langle x(t)|w(t)\rangle > 0 \\ \arctan\left(\dfrac{\langle y(t)|w(t)\rangle}{\langle x(t)|w(t)\rangle}\right) - 180°, & \langle x(t)|w(t)\rangle < 0 \end{cases} \quad (41)$$

$$\tilde{\varphi} = \arctan\left(\frac{\langle z(t)|w(t)\rangle}{\sqrt{\langle x(t)|w(t)\rangle^2 + \langle y(t)|w(t)\rangle^2}}\right) \quad (42)$$

3.2 Estimation dans le domaine temporel

3.2.1 Retard

Pour estimer le retard, on applique la fonction d'intercorrélation normalisée aux signaux $w(t)$ et $a_n(t)$:

$$\chi_{w,a_n}(\tau) = \frac{\langle w|a_n\rangle_\tau}{\|w\| \cdot \|a_n\|_\tau} \quad (43)$$

Dans le cas général avec une source acoustique m le signal de w peut être exprimé à partir des équations (6) et (7):

$$w(t) = \left[g_{m,w} \cdot \delta\left(t - \tau_{m,w}\right) * s_m(t)\right] = g_{m,w} \cdot s_m\left(t - \tau_{m,w}\right) \quad (44)$$

Donc on peut réécrire la fonction d'intercorrélation normalisée comme suit :

$$\chi_{w,a_n}(\tau) = \frac{\langle g_{m,w} \cdot s_m\left(t - \tau_{m,w}\right)\left|a_n(t)\right\rangle_\tau}{\|w\| \cdot \|a_n\|_\tau} \quad (45)$$

où $\tau_{m,w}$ est le retard entre la source et le microphone principal, sous les hypothèses suivantes :

- les trajets indirects et le bruit intrinsèque de mesure sont négligés
- sur une plage temporelle d'observation donnée, une seule source m est active.

Or le signal s_m est relié au signal a_n par l'équation (8), sous les mêmes hypothèses:

$$a_n(t) = g_{m,n} \cdot s_m\left(t - \tau_{m,n}\right) \quad (46)$$

On en déduit donc s_m en fonction de a_n :

$$s_m(t) = \frac{1}{g_{m,n}} a_n\left(t + \tau_{m,n}\right) \quad (47)$$

Cette équation peut également s'écrire de la manière suivante :

$$s_m\left(t - \tau_{m,w}\right) = \frac{1}{g_{m,n}} a_n\left(t + \tau_{m,n} - \tau_{m,w}\right) \quad (48)$$

Il en résulte que l'équation (43) peut s'écrire :

$$\frac{\langle w|a_n\rangle_\tau}{\|w\| \cdot \|a_n\|_\tau} = \frac{\left\langle \frac{g_{m,w}}{g_{m,n}} a_n\left(t + \tau_{m,n} - \tau_{m,w}\right)\middle| a_n(t)\right\rangle_\tau}{\|w\| \cdot \|a_n\|_\tau} \quad (49)$$

Or, en posant $g_{m,n,w} = \frac{g_{m,w}}{g_{m,n}}$, et $\tau_{m,n,w} = -(\tau_{m,n} - \tau_{m,w})$, l'équation (49) peut aussi s'écrire à l'aide de l'équation (13):

$$\frac{\langle w|a_n\rangle_\tau}{\|w\| \cdot \|a_n\|_\tau} = \frac{g_{m,n,w} \cdot \tau_{m,n,w}\langle a_n|a_n\rangle_\tau}{\|w\| \cdot \|a_n\|_\tau} \quad (50)$$

Il est possible de simplifier encore cette équation en exprimant la norme de w grâce aux équations (46) puis (48), et en profitant des notations proposées, d'où :

$$\|w(t)\| = \left\|g_{m,n,w}.\, a_n\left(t - \tau_{m,n,w}\right)\right\| \quad (51)$$

Il en résulte que l'équation (50) peut s'exprimer de la manière suivante :

$$\frac{\langle w|a_n\rangle_\tau}{\|w\|\cdot\|a_n\|_\tau} = \frac{g_{m,n,w}\cdot\tau_{m,n,W}\langle a_n|a_n\rangle_\tau}{\left\|g_{m,n,w}.a_n(t-\tau_{m,n,w})\right\|\cdot\|a_n\|_\tau} \quad (52)$$

En considérant les gains représentés par $g_{m,n,w}$ comme positifs, cette équation se simplifie de la manière suivante :

$$\frac{\langle w|a_n\rangle_\tau}{\|w\|\cdot\|a_n\|_\tau} = \frac{\tau_{m,n,w}\langle a_n|a_n\rangle_\tau}{\|a_n\|_{\tau_{m,n,w}}\cdot\|a_n\|_\tau} \quad (53)$$

On remarque que le second membre de l'équation (53) correspond à la fonction d'intercorrélation normalisée entre le signal $a_n(t-\tau_{m,n,w})$ et le signal $a_n(t)$. Il en résulte que lorsque $\tau = \tau_{m,n,W}$ la fonction (53) donne une valeur maximale unitaire.

Ainsi, pour trouver la valeur recherchée $\tau_{m,n,w}$, il suffit d'identifier la valeur τ pour laquelle l'intercorrélation normalisée entre les signaux connus $w(t)$ et $a_n(t)$ est maximale. Dans le cas d'usage général, les signaux de plusieurs sources sont présents dans le signal principal $w(t)$ alors que le signal appoint $a_n(t)$ est beaucoup plus représentatif de la source sonore dont on veut estimer les paramètres (surtout en l'absence de diaphonie). Il est donc plus judicieux de prendre un morceau du signal d'appoint comme référence et de rechercher dans le signal principal $w(t)$ avec quel décalage temporel on trouve le morceau de signal qui lui ressemble le plus. Autrement dit, on préconise de considérer comme fonction d'intercorrélation normalisée :

$$C(\tau) = \frac{\langle a_n|w\rangle_{-\tau}}{\|a_n\|.\|w\|_{-\tau}} \quad (54)$$

En pratique c'est a priori le signal $w(t)$ qui est en retard sur le signal d'appoint $a_n(t)$. Il s'agit donc généralement de faire une recherche dans $w(t)$ sur une portion de signal plus récente que la portion de signal $a_n(t)$ prise comme référence.

On introduit l'estimateur $\tilde{\tau}$ (mais aussi $\tilde{\theta}$, $\tilde{\varphi}$, $\tilde{g}$) associé au paramètre recherché $\tau_{m,n,w}$ (et respectivement θ_n, φ_n, $g_{m,n,w}$). On définit le retard cible estimé comme le maximum de la fonction d'intercorrélation normalisée de l'équation (54) :

$$\tilde{\tau} = \underset{\tau}{\operatorname{argmax}}(C(\tau)) \quad (55)$$

Dans le cas de plusieurs sources acoustiques l'interférence peut perturber l'estimation (55) et élever des pics secondaires de l'intercorrélation $C(\tau)$ par rapport au pic principal

correspondant au retard ciblé. Pour améliorer l'estimation du retard on peut lisser la fonction d'intercorrélation au cours du temps à l'aide d'un filtre FIR :

$$C_i'(\tau) = \sum_{k=0}^{K-1} b_k C_{i-k}(\tau) \tag{56}$$

où K est l'ordre du filtre, b_k sont les coefficients de filtre et i indique la trame du signal considérée. Si on désire réaliser une fonction de moyennage, on fixe les coefficients de la manière suivante : $b_k = \frac{1}{K+1}$. Cette méthode est très efficace lorsqu'un signal quasi-stationnaire varie en fréquences au cours de temps puisque le pic principal reste stable au cours des accumulations des trames dans la fonction (56) tandis que les pics secondaires varient en position et leurs amplitudes diminuent donc avec chaque trame i. Une autre méthode d'utilisation d'un filtre FIR pour trouver le retard entre les deux signaux peut être trouvée dans un travail [25]. Un exemple de la fonction moyenne basée sur l'équation (56) est présenté sur la Figure 30.

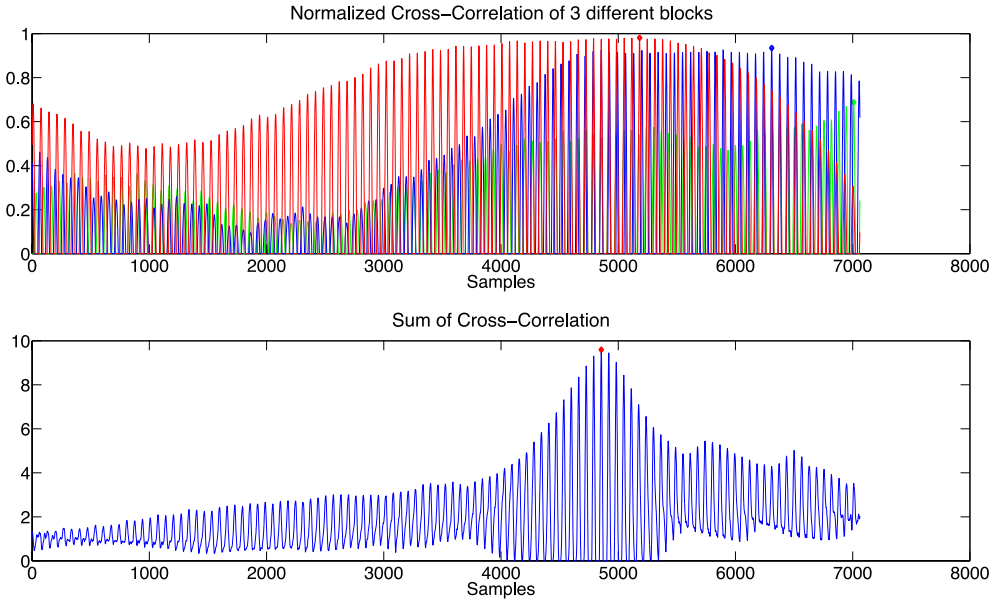

Figure 30. Exemple d'un signal harmonique. Corrélation de trois trames différentes (haut). Intercorrélation lissée (bas).

Une autre manière d'améliorer l'estimation du retard est de lisser la fonction d'intercorrélation (54) à l'aide d'un facteur d'oubli α :

$$C_i'(\tau) = \alpha \cdot C_{i-1}'(\tau) + (1-\alpha) \cdot C_i(\tau) \tag{57}$$

Le facteur d'oubli α peut être fixe ou adaptatif en fonction du signal, mais ce dernier cas n'a pas été étudié. Cette technique a fait l'objet d'une demande de brevet [2].

3.2.2 Position et gain

Les produits scalaires entre les composantes du micro principal (HOA) et les microphones d'appoint s'écrivent comme suit :

$$\begin{cases} \langle x|a_n\rangle_{\tilde{\tau}_n} = \eta \cdot \langle w|a_n\rangle_{\tilde{\tau}_n} \cdot cos\theta_n \cdot cos\varphi_n \\ \langle y|a_n\rangle_{\tilde{\tau}_n} = \eta \cdot \langle w|a_n\rangle_{\tilde{\tau}_n} \cdot sin\theta_n \cdot cos\varphi_n \\ \langle z|a_n\rangle_{\tilde{\tau}_n} = \eta \cdot \langle w|a_n\rangle_{\tilde{\tau}_n} \cdot sin\varphi_n \end{cases} \quad (58)$$

Pour calculer l'azimut θ_n et l'élévation φ_n du signal capté par le microphone d'appoint a_n situé à proximité de la source acoustique, on se place sous les mêmes hypothèses que précédemment.

Le rapport entre la deuxième et la première équation du système (58) permet d'obtenir l'azimut $\tilde{\theta}$ à travers la fonction atan2 :

$$\tilde{\theta}_n = \text{atan2}\left(\langle y|a_n\rangle_{\tilde{\tau}_n}, \langle x|a_n\rangle_{\tilde{\tau}_n}\right) \quad (59)$$

La fonction atan2 présente l'avantage de fournir des mesures d'angles comprises dans un intervalle $[-\pi, \pi]$ alors que la fonction arctangente classique ne permet d'obtenir les angles que dans un intervalle $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$, ce qui laisse une ambiguïté sur des angles diamétralement opposés. On déduit l'élévation $\tilde{\varphi}$ de la dernière équation du système (58):

$$\langle z|a_n\rangle_{\tilde{\tau}_n} = \eta \cdot \langle w|a_n\rangle_{\tilde{\tau}_n} \cdot \sin\tilde{\varphi}_n \quad \Longrightarrow \quad \tilde{\varphi}_n = \arcsin\left(\frac{\langle z|a_n\rangle_{\tilde{\tau}_n}}{\eta \cdot \langle w|a_n\rangle_{\tilde{\tau}_n}}\right) \quad (60)$$

A partir de l'estimateur $\tilde{\tau}_n$ donné par l'équation (55) le niveau du gain $\tilde{g}_{m,n,w}$ peut être estimé comme un rapport entre le produit scalaire du signal de microphone principal et du signal de microphone d'appoint, et le produit scalaire du signal de microphone d'appoint par lui-même :

$$\tilde{g}_{m,n,w} = \frac{\langle w|a_n\rangle_{\tilde{\tau}_n}}{{}_{\tilde{\tau}_n}\langle a_n|a_n\rangle_{\tilde{\tau}_n}} \quad (61)$$

Cette estimation du gain est celle qui minimise la différence entre le signal a_n et le signal w, sur la fenêtre temporelle considérée, au sens des moindres carrés.

On notera que les estimateurs ci-dessus sont déterminés en appliquant un retard au signal d'appoint $a_n(t)$ alors que formellement, c'est au signal principal $w(t)$ que l'on applique un retard opposé (donc une avance) pour la recherche dudit retard. Ces estimateurs restent valables en considérant qu'ils s'appliquent avec un décalage temporel supplémentaire commun aux deux signaux. En rectifiant cet aspect, on obtient finalement tous les paramètres qui permettent de retarder, spatialiser et mixer le microphone d'appoint au microphone principal :

$$\tilde{\tau}_n = \underset{\tau}{\text{Argmax}}\, C_i\,(\tau) \tag{62}$$

$$\tilde{\theta}_n = \text{atan2}\big(\langle a_n|y\rangle_{-\tilde{\tau}_n}, \langle a_n|x\rangle_{-\tilde{\tau}_n}\big) \tag{63}$$

$$\tilde{\varphi}_n = \arcsin\left(\frac{\langle a_n|z\rangle_{-\tilde{\tau}_n}}{\eta \cdot \langle a_n|w\rangle_{-\tilde{\tau}_n}}\right) \tag{64}$$

$$\tilde{g}_{m,n,w} = \frac{\langle a_n|w\rangle_{-\tilde{\tau}_n}}{\|a_n\|^2} \tag{65}$$

3.3 Estimation dans le domaine fréquentiel

L'analyse dans le domaine fréquentiel peut permettre d'améliorer l'estimation des paramètres de sources acoustiques. Dans le cas d'une source sonore unique, l'estimation des paramètres est relativement facile dans la mesure où les signaux mesurés par le microphone principal résultent de cette seule source. En revanche, dans le cas où plusieurs sources sont présentes, les signaux mesurés par le microphone principal mélangent les contributions provenant des différentes sources. Le spectre d'une source sonore peut alors interférer avec celui d'une autre source sonore (Figure 31). Ici on parle de la fréquence *interférée*, ça veut dire que la fréquence est partagée par au moins deux sources acoustiques. Cependant, il est aussi possible qu'il existe des bandes de fréquences pour lesquelles une seule source est prédominante (*non-interférée*). L'analyse des signaux dans une de ces bandes permet alors une estimation robuste. Dans un exemple présenté sur la Figure 31, les fréquences interférées et non-interférées sont choisies à partir du rapport (en dB) des amplitudes spectrales de chaque source acoustique. Si le rapport dépasse un certain seuil (arbitraire ou choisi manuellement par l'ingénieur du son), il est donc possible de localiser une source prédominante.

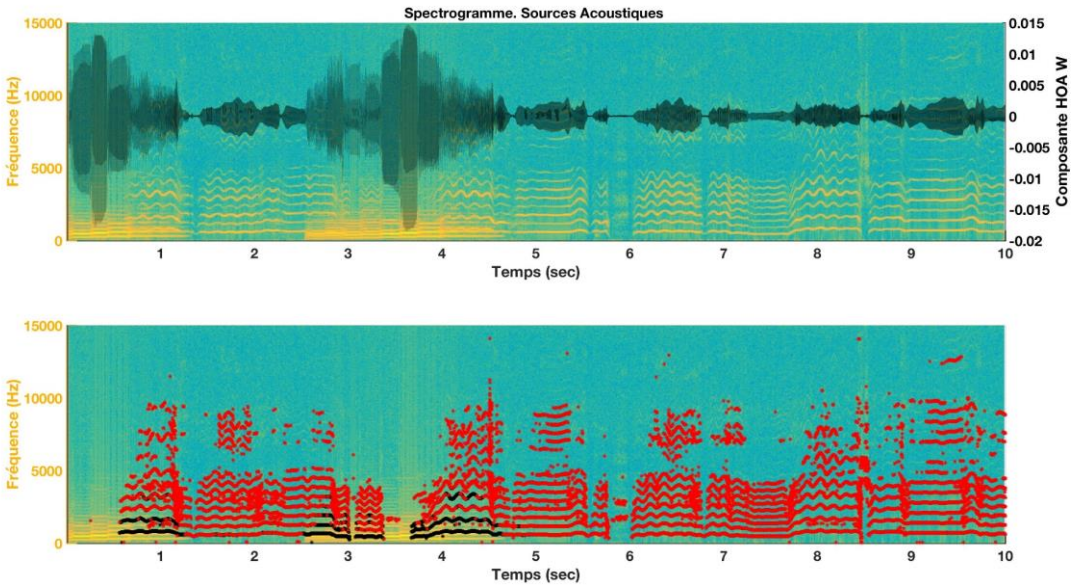

Figure 31. Deux sources acoustiques avec les fréquences interférées (points noirs) et non-interférées (point rouges).

Les signaux des microphones d'appoint $a_n(t)$, des sources acoustiques $s_m(t)$ (Figure 22) et les composantes d'Ambisonics d'ordre 1 : $w(t), x(t), y(t), z(t)$ peuvent être représentés dans le domaine fréquentiel à un instant donné à travers la transformée de Fourier à court-terme : $A_n(\omega)$, $S_m(\omega)$ et $W(\omega)$, $X(\omega)$, $Y(\omega)$, $Z(\omega)$ où ω est un bin fréquentiel.

On considère les mêmes équations (63), (64) de l'estimation de la position que dans le domaine temporel. En prenant en compte la version fréquentielle de tous les membres des équations (63), (64) on peut les exprimer pour le domaine fréquentiel :

$$\tilde{\theta}_n = \text{atan2}(\langle A_n(\omega)|Y(\omega)\rangle_{-\tilde{\tau}}, \langle A_n(\omega)|X(\omega)\rangle_{-\tilde{\tau}}) \tag{66}$$

$$\tilde{\varphi}_n = \arcsin\left(\frac{\langle A_n(\omega)|Z(\omega)\rangle_{-\tilde{\tau}}}{\eta \cdot \langle A_n(\omega)|W(\omega)\rangle_{-\tilde{\tau}}}\right) \tag{67}$$

où l'indice $-\tilde{\tau}$ dans chaque produit scalaire représente le décalage temporel du signal de cette valeur avant d'être transformé dans le domaine fréquentiel.

Dans les équations (66) et (67), contrairement au domaine temporel, on opère avec des grandeurs complexes. On peut montrer que le lien entre les produits scalaires dans le domaine temporel et fréquentiel s'exprime comme suit :

$$\langle s_1(t)|s_2(t)\rangle = \mathfrak{R}(\langle S_1(\omega)|S_2^*(\omega)\rangle) \tag{68}$$

A un instant donné, et pour un bin fréquentiel ω_q, le produit scalaire (68) s'exprime comme suit :

$$\mathfrak{R}\left(\langle S_1(\omega_q)|S_2^*(\omega_q)\rangle\right) = \mathfrak{R}\left(S_1(\omega_q) \cdot S_2(\omega_q)\right) \tag{69}$$

Donc avec (69) les équations (66) et (67) s'écrivent sous la forme suivante :

$$\tilde{\theta}_n = \operatorname{atan}\left(\mathfrak{R}\left(\frac{A_n(\omega_q)Y(\omega_q)_{-\tilde{\tau}}}{A_n(\omega_q)X(\omega_q)_{-\tilde{\tau}}}\right)\right) = \operatorname{atan}\left(\mathfrak{R}\left(\frac{Y(\omega_q)_{-\tilde{\tau}}}{X(\omega_q)_{-\tilde{\tau}}}\right)\right) \quad (70)$$

$$\tilde{\varphi}_n = \arcsin\left(\mathfrak{R}\left(\frac{Z(\omega_q)_{-\tilde{\tau}}}{\eta \cdot W(\omega_q)_{-\tilde{\tau}}}\right)\right) \quad (71)$$

où $W(\omega_q)_{-\tilde{\tau}}$, $X(\omega_q)_{-\tilde{\tau}}$, $Y(\omega_q)_{-\tilde{\tau}}$ et $Z(\omega_q)_{-\tilde{\tau}}$ sont les composantes HOA du domaine temporel retardées de $\tilde{\tau}$ et transformées dans le domaine fréquentiel. A partir des équations (70) et (71) on peut noter que pour un bin fréquentiel la position angulaire ne prend pas en compte le signal du microphone d'appoint.

3.4 Descripteurs et indice de confiances

On a considéré l'estimation des paramètres dans les domaines temporel et fréquentiel. En revanche, la question de l'efficacité de l'estimation reste ouverte pour le cas d'une scène sonore complexe, c'est-à-dire en présence de bruit et de réverbération et/ou en présence de plusieurs sources acoustiques qui peuvent interférer les unes avec les autres.

On peut estimer le retard avec (62) et l'azimut et l'élévation soit avec (63), (64) soit avec (70), (71) mais il peut être intéressant d'évaluer la qualité de l'estimation. En particulier, on peut souhaiter connaître ou prédire la précision de l'estimation du retard, ainsi que la précision de la localisation.

Pour répondre à cette problématique on peut utiliser des indicateurs (ou « descripteurs »), c'est-à-dire des fonctions permettant de prédire l'efficacité de l'estimation des paramètres à partir des caractéristiques du champ sonore. Un premier exemple de ce type de descripteurs est l'énergie du signal de microphone d'appoint. Le signal du microphone d'appoint est plus net que le signal capté par le microphone principal puisque ce dernier capte toute la scène sonore avec des sources acoustiques mélangées. Donc l'énergie du signal du microphone d'appoint peut indiquer les endroits significatifs pour l'analyse et prédire dans quels endroits l'estimation sera efficace. On peut exprimer l'énergie par rapport à l'énergie de référence E_0 (par exemple, un signal carré avec énergie maximale de 1):

$$E_{a_n} = 10 \cdot \log_{10}\left(\frac{\frac{1}{K-1}\sum_{k=0}^{K-1} a(k)^2}{E_0}\right) \quad (72)$$

Avec le descripteur (72) on peut sélectionner les endroits significatifs pour l'analyse. Une utilisation possible de ce descripteur est illustrée sur la Figure 32. Dans cet exemple on observe la valeur de E_{a_n} et on considère comme signal utile les segments pour lesquels cette valeur est supérieur au seuil de -60 dB (arbitraire).

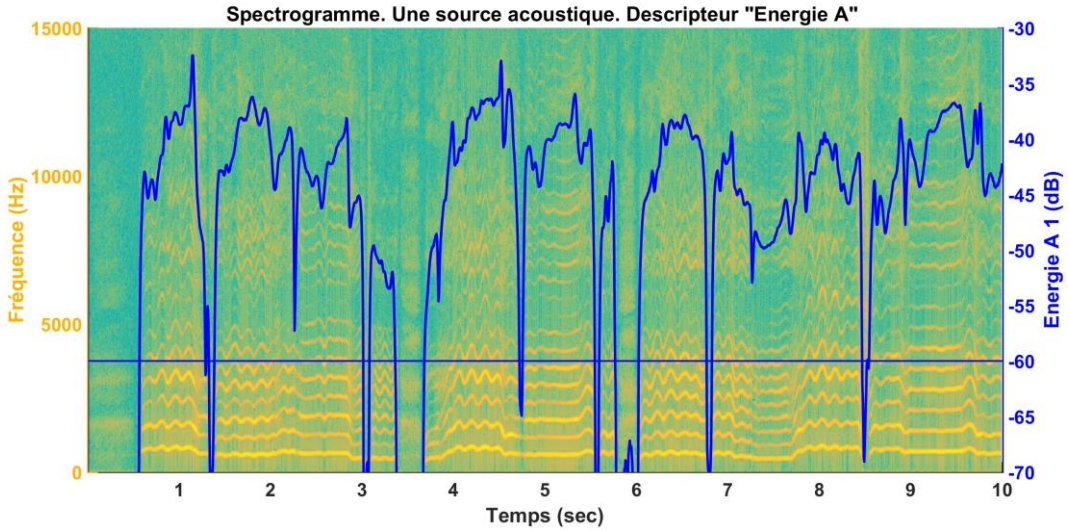

Figure 32. Source acoustique (voix) et le descripteur E_{a_n}. La sélection d'un signal utile à partir d'un taux de -60 dB.

Des descripteurs peuvent également être utiles pour la recherche du retard par la fonction d'intercorrélation. Dans le cas d'un signal périodique la fonction d'intercorrélation risque de présenter plusieurs pics. En présence de bruit, la sélection de la valeur maximale peut donc aboutir à une erreur sur la valeur de retard. On note également qu'en présence d'une attaque ou d'un « transitoire » selon un terme consacré du domaine du traitement du signal, la fonction d'intercorrélation ne présente généralement qu'un seul pic bien distinct. Un descripteur pertinent consiste donc à mesurer la différence d'amplitude entre les 2 pics principaux de la fonction d'intercorrélation. Cette fonction fournit une information robuste (plus robuste que la valeur maximale de l'intercorrélation, qui peut être maximale dans le cas d'un signal périodique) sur le niveau de confiance à accorder à l'estimateur du retard.

On définit le retard correspondant au premier pic principal « I » de la fonction d'intercorrélation par analogie avec (62) :

$$\tilde{\tau}_{max\mathrm{I}} = \underset{\tau}{\mathrm{Argmax}}\, C(\tau) \tag{73}$$

Donc le retard associé au secondaire pic « II » peut être exprimé comme suit :

$$\tilde{\tau}_{max\mathrm{II}} = \underset{\tau \neq \tilde{\tau}_{max\mathrm{I}}}{\mathrm{Argmax}}\, C(\tau) \tag{74}$$

où $C(\tau)$ est une intercorrélation entre les deux signaux (54), (56) ou (57).

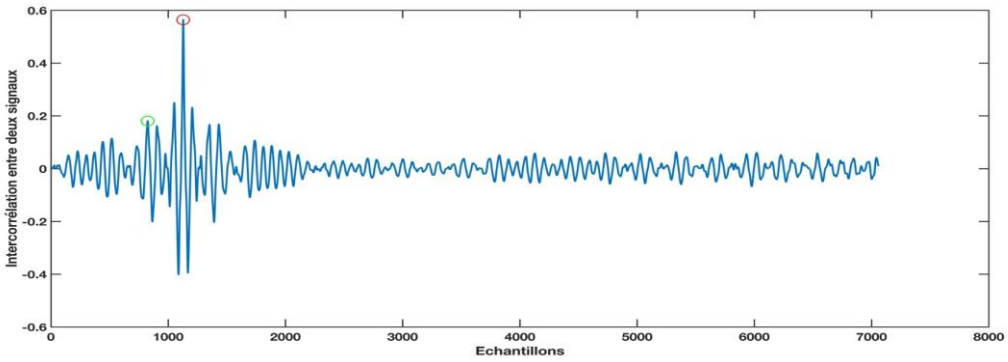

Figure 33. Intercorrélation entre deux signaux avec les pics principal (rouge) et secondaire (vert).

Afin de ne pas prendre en compte les voisins de la valeur maximale de l'intercorrélation qui appartiennent au même pic (ce qui correspond à la décroissance naturelle de la fonction d'intercorrélation), il est nécessaire d'exclure un certain voisinage. Il existe plusieurs méthodes de détection des pics secondaires dans le domaine temporel qui sont bien décrites dans [26]. Par exemple, on peut exclure toutes les valeurs successives voisines inférieures à 5% de la valeur maximale, ou bien ne considérer un pic secondaire que lorsque la valeur de la fonction d'intercorrélation est descendue, entre le pic principal et le pic secondaire considéré, sous un certain seuil relatif à la valeur maximale. Ce seuil peut être zéro, auquel cas le critère considéré est le changement de signe de la fonction d'intercorrélation entre deux pics retenus.

Après avoir détecté les pics principal et secondaire (voir Figure 33) par une des méthodes décrites dans[26], avec (73) et (74) on peut calculer le descripteur suivant, correspondant au rapport des amplitudes des pics principal et secondaire :

$$R_{\mathrm{II/I}} = \frac{C(\tilde{\tau}_{max\mathrm{II}})}{C(\tilde{\tau}_{max\mathrm{I}})} \qquad (75)$$

Dans le cas d'un signal périodique avec (75) on obtient la valeur $R_{\mathrm{II/I}}$ qui converge vers « 1 » car les amplitudes de deux pics maximaux se sont relativement proches. A l'inverse, dans le cas d'un signal transitoire la valeur $R_{\mathrm{II/I}}$ sera minimale (Figure 34).

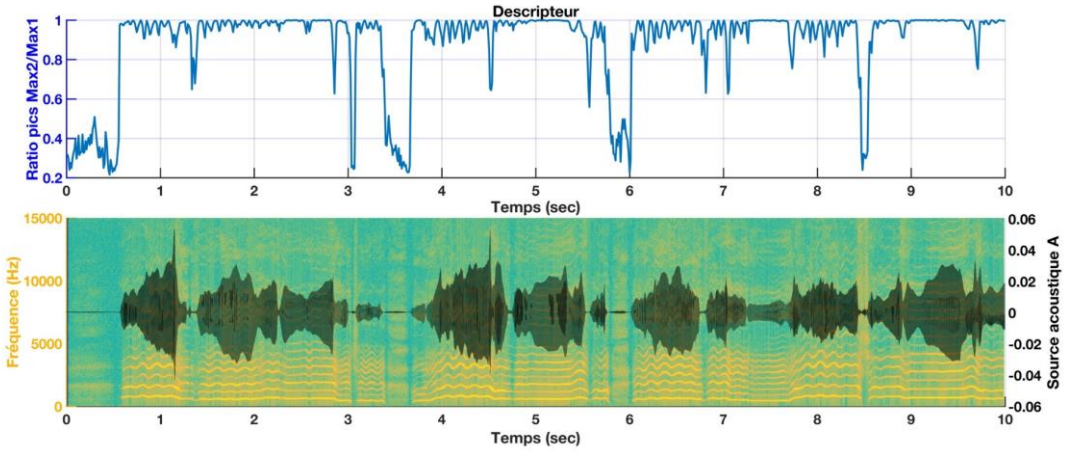

Figure 34. Un signal périodique avec le descripteur $R_{II/I}$ qui est proche de 1 au cas avec la présence du signal périodique.

Un autre descripteur possible est donnée par la valeur de l'amplitude $C(\tilde{\tau})$ où $\tilde{\tau}$ est estimé par (55) ou (62). Plus cette valeur est élevée, et plus les deux signaux comparés se ressemblent, ce qui suggère la présence d'une source unique ou l'absence de bruit ou de réverbération.

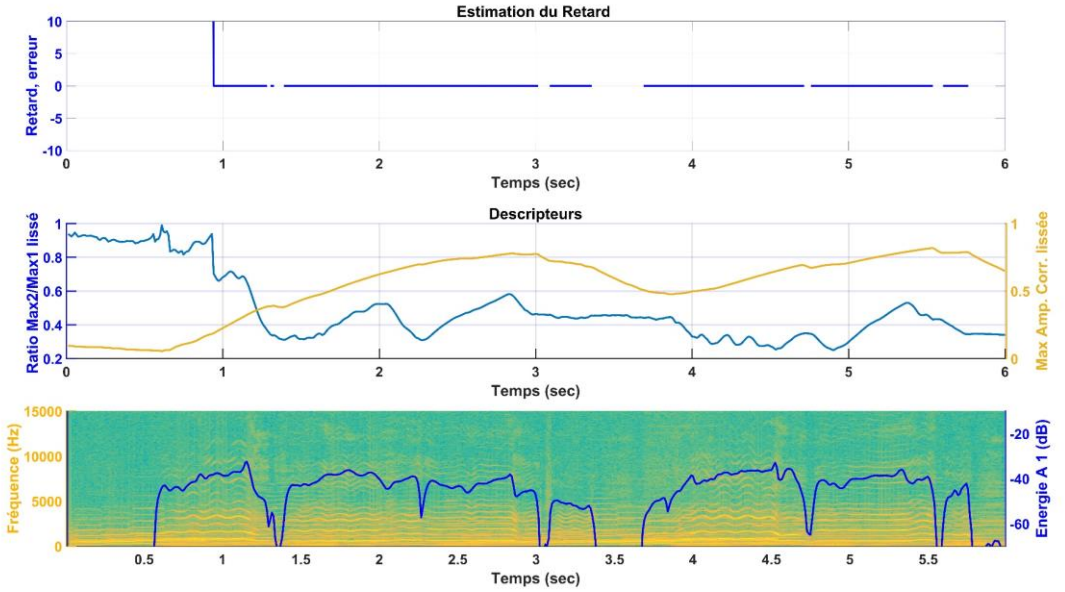

Figure 35. Estimation du retard d'une source acoustique dans la scène sonore composée de plusieurs sources acoustiques.

Les descripteurs (72), (75) ainsi que l'amplitude $C(\tilde{\tau})$ peuvent améliorer l'estimation du retard (voir un exemple sur la Figure 35). Dans le chapitre 3.2.2 on a évoqué le fait que pour trouver la position de la source acoustique il fallait tout d'abord décaler le signal du microphone principal (composante $w(t)$) par le retard estimé $\tilde{\tau}$ avant d'appliquer les équations. On dispose alors du signal du microphone d'appoint $a(t)$ et des composantes HOA qui sont en retard de $\tilde{\tau}$

sur le signal de microphone d'appoint: $w(t-\tilde{\tau})$, $x(t-\tilde{\tau})$, $y(t-\tilde{\tau})$ et $z(t-\tilde{\tau})$. Les composantes HOA contiennent toutes les sources acoustiques captées par le microphone principal avec les niveaux de gain associés. On peut donc déterminer la contribution de chacune des sources aux signaux mesurés par le microphone principal en calculant le rapport d'énergie du signal $a(t)$ et de la première composante HOA par exemple $w(t-\tilde{\tau})$ (en dB) :

$$E_{a/w} = \log_{10}\left(\frac{{}^{1}/_{K}\sum_{k=0}^{K-1} a(k)^2}{E_0}\right) - \log_{10}\left(\frac{{}^{1}/_{K}\sum_{k=0}^{K-1} w(k-d)^2}{E_0}\right) \quad (76)$$

où $d = \tilde{\tau} \cdot f_s$ avec une fréquence d'échantillonnage f_s, K représente la taille des blocs d'échantillons considérés et $k = t \cdot f_s$. La valeur E_0 est une valeur d'énergie de référence. La valeur $E_{a/w}$ peut être utilisée comme un descripteur indiquant les segments pour lesquels la source correspondant au microphone d'appoint contribue significativement aux signaux mesurés par microphone principal (Figure 36).

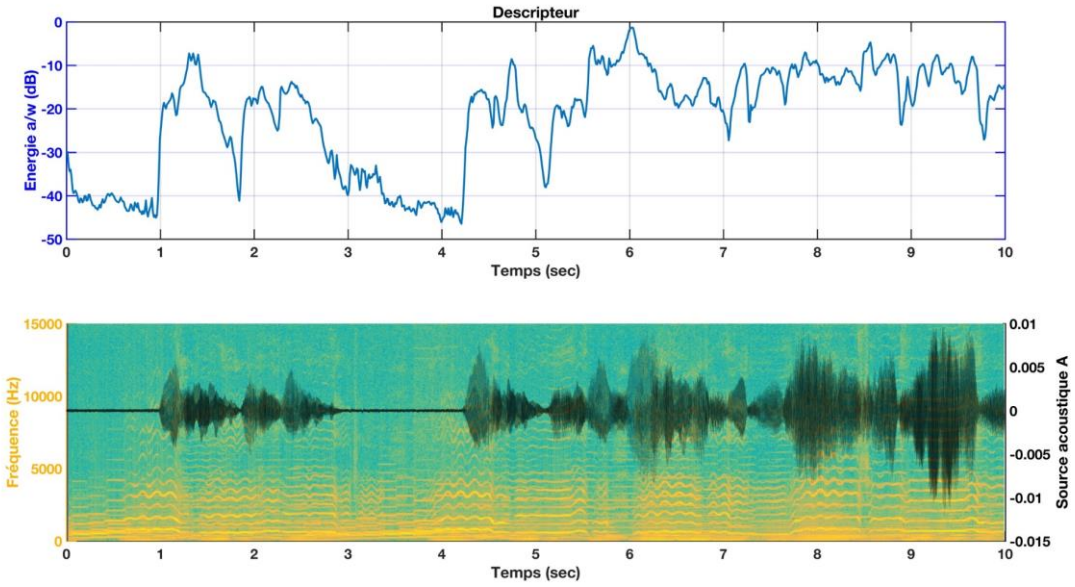

Figure 36. Contribution d'un signal dans le signal du microphone principal représentée par le descripteur $E_{a/w}$. La scène sonore est composé par 7 sources acoustiques (instruments dans une pièce de Mozart).

En résumé, on a considéré quatre descripteurs qui peuvent à aider prédire la qualité de l'estimation des paramètres :

$$E_{a_n} = 10 \cdot \log_{10}\left(\frac{{}^{1}/_{K}\sum_{k=0}^{K} a(k)^2}{E_0}\right) \quad (77)$$

$$C(\tilde{\tau}) = \frac{\langle a_n | w \rangle_{-\tilde{\tau}}}{\|a_n\| . \|w\|_{-\tilde{\tau}}} \tag{78}$$

$$R_{\mathrm{II/I}} = \frac{C(\tilde{\tau}_{max\mathrm{II}})}{C(\tilde{\tau}_{max\mathrm{I}})} \tag{79}$$

$$E_{a/w} = \frac{\log_{10}\left(\frac{1/K \sum_{k=0}^{K} a(k)^2}{E_0}\right)}{\log_{10}\left(\frac{1/K \sum_{k=0}^{K} w(k-d)^2}{E_0}\right)} \tag{80}$$

Les descripteurs (78) et (79) peuvent être également exprimés avec la fonction d'intercorrélation (56) ou (57) pour le signal qui est divisé par des trames de taille fixe (voir chapitre 4.1).

A partir des descripteurs (77) - (80) on peut composer un indice de confiance sur l'estimation des paramètres. L'indice de confiance peut agréger plusieurs descripteurs, l'idée étant de fournir une note correspondant à la qualité de l'estimation effectuée. Cette note peut aider à déterminer si l'estimation est « bonne » (à conserver) ou « mauvaise » (à jeter).

On peut par exemple composer un indice de confiance valable pour l'analyse « trame par trame » avec (56) et (79) qui est associé à l'estimation du retard :

$$IC_1 = R_{\mathrm{II/I}} \cdot C'(\tilde{\tau}) \tag{81}$$

où $C'(\tilde{\tau})$ est une fonction intercorrélation obtenue par l'équation (56) ou (57). L'idée principale de cet indice de confiance est d'utiliser d'un côté le descripteur $C'(\tilde{\tau})$ qui mémorise l'information sur l'estimation des trames précédentes et de l'autre côté $R_{\mathrm{II/I}}$ qui peut augmenter la valeur de l'indice dans le cas d'un signal transitoire (Figure 37). Dans le cas d'un signal périodique dont la fonction d'intercorrélation présente deux pics relativement proches, $C'(\tilde{\tau})$ accumule la contribution des intercorrélations précédentes mais $R_{\mathrm{II/I}}$ a une valeur minimale. Dans cette situation l'indice de confiance $IC_{\tilde{\tau}}$ indique que l'estimation est moins robuste que dans le cas de l'estimation d'un signal transitoire.

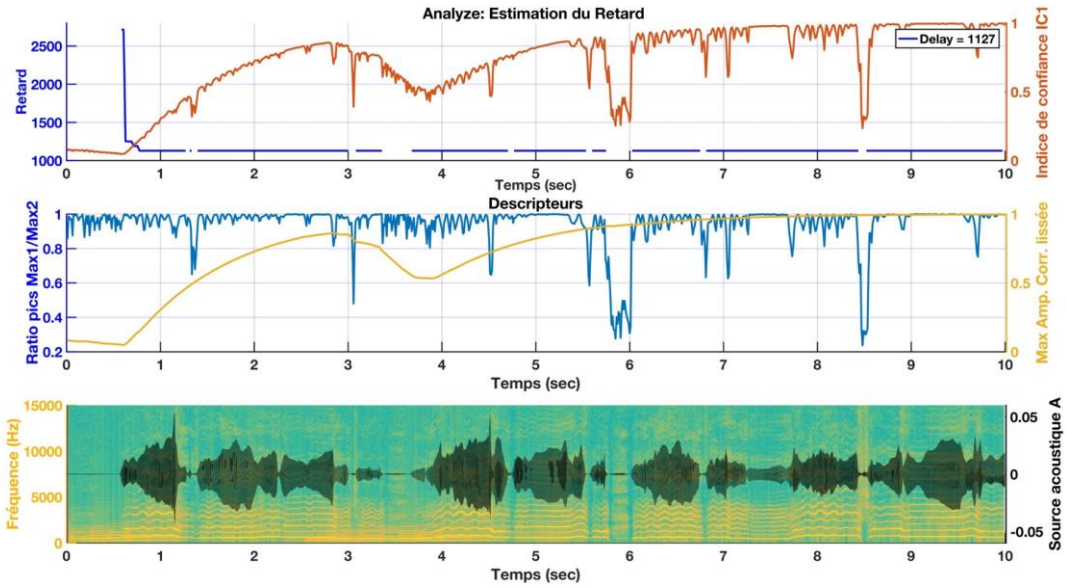

Figure 37. Exemple d'estimation du retard d'un signal avec son indice de confiance IC_1 dans une scène sonore composée par deux sources acoustiques. Haut : le retard estimé avec l'indice de confiance IC_1 ; milieu : deux descripteurs qui composent l'indice de confiance IC_1 ; bas : signal acoustique et son spectrogramme.

Un autre indice de confiance pouvant être proposé s'appuie sur le descripteur $C'(\tilde{\tau})$ et le rapport d'énergies $E_{a/w}$ (80) :

$$IC_2 = C'(\tilde{\tau}) \cdot E_{a/w} \tag{82}$$

Ici les valeurs accumulées par la fonction $C'(\tilde{\tau})$ sont multipliées par le rapport d'énergies entre le microphone d'appoint et la première composante HOA w d'une trame donnée. Cet indice de confiance est donc plus élevé dans les endroits (trames) où le signal capté par le microphone d'appoint est plus présent dans le signal capté par le microphone principal (Figure 38).

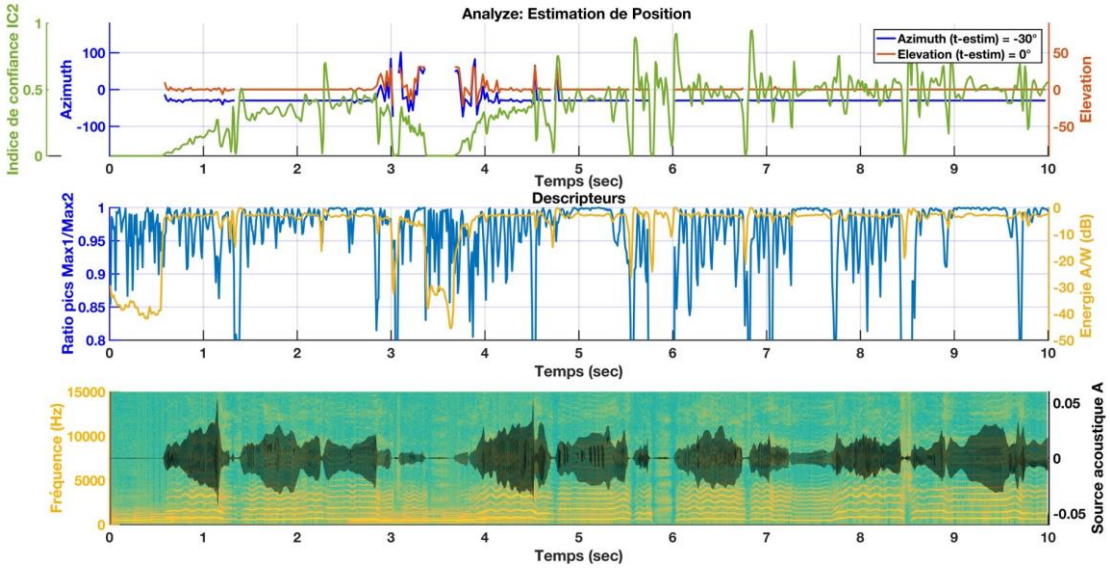

Figure 38. Exemple d'estimation de l'azimut et de l'élévation d'un signal avec son indice de confiance IC_2 dans une scène sonore composée par deux sources acoustiques. Haut : l'azimut et l'élévation estimés avec l'indice de confiance IC_2 ; milieu : deux descripteurs qui composent l'indice de confiance IC_2 ; bas : signal acoustique et son spectrogramme.

L'indice de confiance IC_1 avec le facteur d'oubli α (57) a été intégré dans l'algorithme présenté dans le chapitre 4 pour donner une note à l'estimation du retard. L'indice IC_2 est lui utilisé pour l'estimation de la position des sources acoustiques. Bien entendu, d'autres indices de confiance peuvent être proposés à partir de différents descripteurs. Dans la section suivante on s'intéresse de plus près à la notion de vecteur vélocité qui peut aussi être utilisé pour prédire la précision de l'estimation des paramètres de mixage.

Vecteur vélocité

Dans le chapitre 3.4 on a déjà introduit le vecteur vélocité qui joue un rôle dans l'estimation des paramètres de source acoustique par la méthode DirAC. Dans cette partie on utilise la norme de vecteur vélocité en prenant en compte le signal fourni par le microphone d'appoint a_n. Les composantes du vecteur vélocité sont données par :

$$v_x = \frac{\langle a_n(t)|x(t)\rangle_{-\tilde{\tau}}}{\langle a_n(t)|w(t)\rangle_{-\tilde{\tau}}}; \quad v_y = \frac{\langle a_n(t)|y(t)\rangle_{-\tilde{\tau}}}{\langle a_n(t)|w(t)\rangle_{-\tilde{\tau}}}; \quad v_z = \frac{\langle a_n(t)|z(t)\rangle_{-\tilde{\tau}}}{\langle a_n(t)|w(t)\rangle_{-\tilde{\tau}}}. \tag{83}$$

On en déduit la norme du vecteur vélocité :

$$r_v = \sqrt{v_x^2 + v_y^2 + v_z^2} \tag{84}$$

On peut montrer le lien entre les équations d'estimation (63), (64) et les composantes du vecteur vélocité (83). Ainsi les paramètres angulaire peuvent être exprimés en fonction de v_x, v_y et v_z:

$$\tilde{\theta}_n = \operatorname{atan}\left(\frac{v_y}{v_x}\right) \tag{85}$$

$$\tilde{\varphi}_n = \arcsin\left(\frac{1}{\eta} v_z\right) \tag{86}$$

La norme du vecteur vélocité r_v peut être à la base d'un nouveau descripteur. Une valeur de norme $r_v = 1$ montre que le vecteur vélocité est orienté vers la source. Les composantes v_x, v_y et v_z peuvent être décrites soit dans le domaine temporel (83), (84) soit dans le domaine fréquentiel :

$$V_x = \frac{\langle A_n(\omega)|X(\omega)\rangle_{-\tilde{\tau}}}{\langle A_n(\omega)|W(\omega)\rangle_{-\tilde{\tau}}}; \quad V_y = \frac{\langle A_n(\omega)|Y(\omega)\rangle_{-\tilde{\tau}}}{\langle A_n(\omega)|W(\omega)\rangle_{-\tilde{\tau}}}; \quad V_z = \frac{\langle A_n(\omega)|Z(\omega)\rangle_{-\tilde{\tau}}}{\langle A_n(\omega)|W(\omega)\rangle_{-\tilde{\tau}}};$$
$$r_V = \sqrt{V_x^2 + V_y^2 + V_z^2} \tag{87}$$

Les fonctions $A_n(\omega)$, $X(\omega), Y(\omega)$, $Z(\omega)$ et $W(\omega)$ sont à la base complexe. Donc dans les équations (87) on a le produit scalaire entre les valeurs complexes par analogie avec les équations dans le chapitre 3.3. Donc pour une analyse dans le domaine fréquentiel on s'intéresse à deux descripteurs basés sur la norme du vecteur vélocité : $r_{V_{Im}}$ et $r_{V_{Re}}$ qui sont la norme des parties imaginaire et réelle du vecteur V. La partie réelle ainsi que la norme $r_{V_{Re}}$ permet de prédire l'efficacité de l'estimation de la position de la source par analogie avec r_v dans le domaine temporel. Si $r_{V_{Re}}$ est plus proche de 1 (Figure 39) on peut dire que l'estimation est plutôt fiable. La partie imaginaire $r_{V_{Im}}$, d'autre part, indique la dispersion d'énergie dans le champ acoustique ; les zones de moindre dispersion correspondent aux zones où l'estimation est plus fiable (Figure 39).

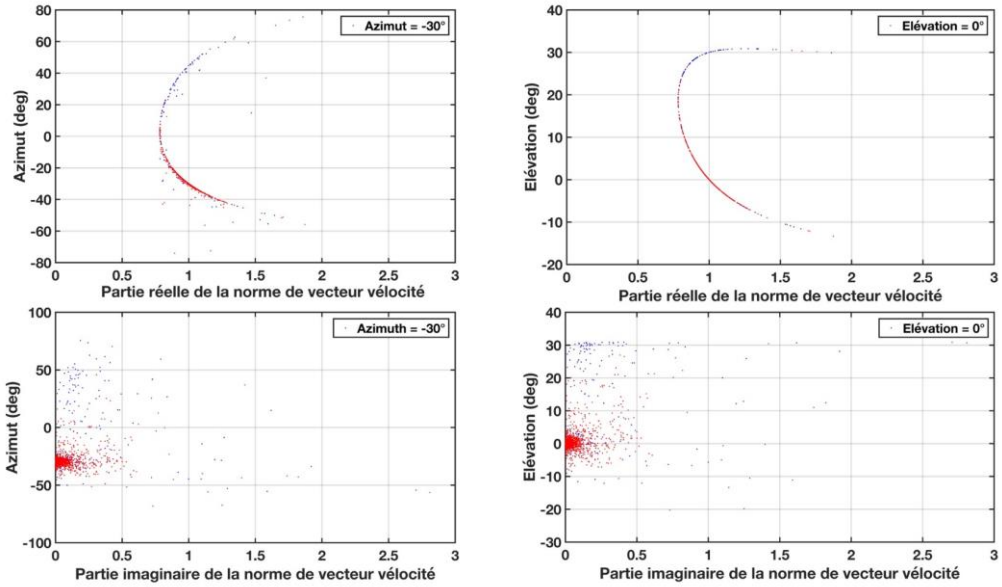

Figure 39. Estimation de la position dans une scène sonore composé par deux sources acoustiques. Le nuage des points de la norme des parties réelle et imaginaire du vecteur vélocité pour les fréquences non-interférées (points rouges) et interférées avec la deuxième source acoustique (points bleus).

Pour calculer $r_{V_{Re}}$ et $r_{V_{Im}}$ on peut utiliser les équations (87) pour toute une bande de fréquence, ou bien bin fréquentiel par bin fréquentiel. Pour un bin fréquentiel ω_q le produit scalaire devient :

$$\mathfrak{R}(V_x) = \mathfrak{R}\left(\frac{\langle A_n(\omega_q)|X(\omega_q)\rangle_{-\tilde{\tau}}}{\langle A_n(\omega_q)|W(\omega_q)\rangle_{-\tilde{\tau}}}\right) = \mathfrak{R}\left(\frac{A_n(\omega_q)X(\omega_q)_{-\tilde{\tau}}}{A_n(\omega_q)W(\omega_q)_{-\tilde{\tau}}}\right) = \mathfrak{R}\left(\frac{X(\omega_q)_{-\tilde{\tau}}}{W(\omega_q)_{-\tilde{\tau}}}\right);$$

$$\mathfrak{R}(V_y) = \mathfrak{R}\left(\frac{Y(\omega_q)_{-\tilde{\tau}}}{W(\omega_q)_{-\tilde{\tau}}}\right) \tag{88}$$

$$\mathfrak{R}(V_z) = \mathfrak{R}\left(\frac{Z(\omega_q)_{-\tilde{\tau}}}{W(\omega_q)_{-\tilde{\tau}}}\right)$$

$$r_{V_{Re}} = \sqrt{(\mathfrak{R}(V_x))^2 + (\mathfrak{R}(V_y))^2 + (\mathfrak{R}(V_z))^2}$$

Par analogie avec (88) on peut en déduire $r_{V_{Im}}$:

$$\begin{aligned}\mathfrak{I}(V_x) &= \mathfrak{I}\left(\frac{X(\omega_q)_{-\tilde{\tau}}}{W(\omega_q)_{-\tilde{\tau}}}\right);\\ \mathfrak{I}(V_y) &= \mathfrak{I}\left(\frac{Y(\omega_q)_{-\tilde{\tau}}}{W(\omega_q)_{-\tilde{\tau}}}\right)\\ \mathfrak{I}(V_z) &= \mathfrak{I}\left(\frac{Z(\omega_q)_{-\tilde{\tau}}}{W(\omega_q)_{-\tilde{\tau}}}\right)\\ r_{V_{Im}} &= \sqrt{(\mathfrak{I}(V_x))^2 + (\mathfrak{I}(V_y))^2 + (\mathfrak{I}(V_z))^2}\end{aligned} \tag{89}$$

Par ailleurs on propose d'utiliser le vecteur de vélocité pour construire une cartographie du champ sonore (voir Annexe 3).

En prenant en compte la notation de [6] et du chapitre 2 le signal encodé en HOA s'exprime :

$$\mathbf{b} = (\boldsymbol{w}\ \boldsymbol{x}\ \boldsymbol{y}\ \boldsymbol{z}\ \ldots)^t \tag{90}$$

La projection de la première composante HOA sur toutes les composantes peut être définie comme suit :

$$\langle \boldsymbol{w}|\mathbf{b}\rangle = [\langle \boldsymbol{w}|\boldsymbol{w}\rangle \quad \langle \boldsymbol{w}|\boldsymbol{x}\rangle \quad \langle \boldsymbol{w}|\boldsymbol{y}\rangle \quad \langle \boldsymbol{w}|\boldsymbol{z}\rangle \quad \ldots] \tag{91}$$

Donc avec le produit scalaire (91) on peut exprimer le vecteur vélocité par analogie avec (83) et (87) :

$$\boldsymbol{v} = \frac{\langle \boldsymbol{w}|\mathbf{b}\rangle}{\langle \boldsymbol{w}|\boldsymbol{w}\rangle} \tag{92}$$

La projection du vecteur d'un signal de microphone d'appoint $a(t)$ sur les composantes HOA s'exprime :

$$\langle \boldsymbol{a}|\mathbf{b}\rangle = [\langle \boldsymbol{a}|\boldsymbol{w}\rangle \quad \langle \boldsymbol{a}|\boldsymbol{x}\rangle \quad \langle \boldsymbol{a}|\boldsymbol{y}\rangle \quad \langle \boldsymbol{a}|\boldsymbol{z}\rangle \quad \ldots] \tag{93}$$

A partir de l'équation (93) on peut exprimer le vecteur vélocité pour le signal d'un microphone d'appoint :

$$\boldsymbol{v}_a = \frac{\langle \boldsymbol{a}|\mathbf{b}\rangle}{\langle \boldsymbol{a}|\boldsymbol{w}\rangle} \tag{94}$$

Dans le cas où plusieurs sources acoustiques sont présentes l'équation d'encodage HOA (90) peut être écrite sous la forme suivante :

$$\mathbf{b} = (\{\boldsymbol{w_1} + \boldsymbol{w_2} + \boldsymbol{w_3} + \ldots\} \quad \{\boldsymbol{x_1} + \boldsymbol{x_2} + \boldsymbol{x_3} + \ldots\} \quad \{\boldsymbol{y_1} + \boldsymbol{y_2} + \boldsymbol{y_3} + \ldots\} \ldots)^t \tag{95}$$

Si les sources acoustiques sont décorrélées le vecteur vélocité pour le champ résultant s'exprime :

$$\mathbf{v} = \frac{\langle\{\mathbf{w_1} + \mathbf{w_2} + \mathbf{w_3} + \ldots\}|\mathbf{b}\rangle}{\langle\mathbf{w_1}|\mathbf{w_1}\rangle + \langle\mathbf{w_2}|\mathbf{w_2}\rangle + \langle\mathbf{w_3}|\mathbf{w_3}\rangle + \ldots} \tag{96}$$

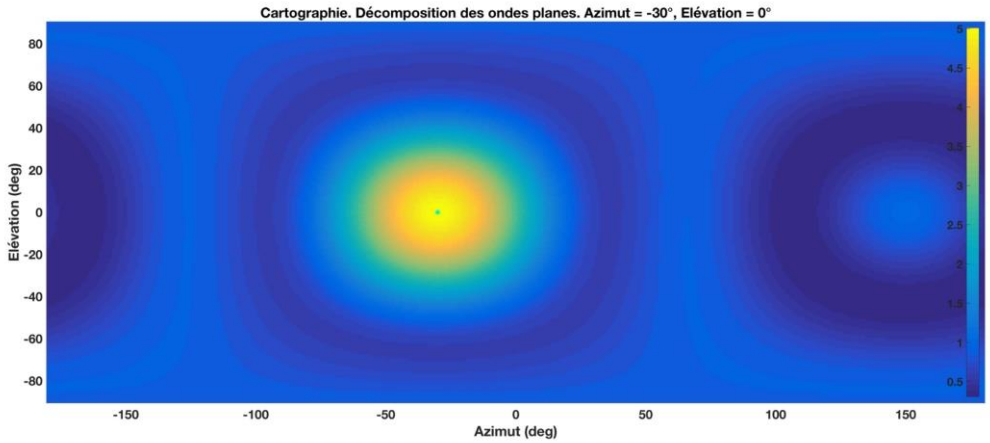

Figure 40. Cartographie d'une source acoustique pour une bande de fréquences.

Le champ sonore avec une source acoustique peut être représenté par sa décomposition en ondes planes dans le plan 2D (Figure 40). Donc la méthode de cartographie est un autre moyen d'étudier la position de la source acoustique en espace par le vecteur vélocité. En particulier, sur la Figure 40 le point vert dans un endroit plus clair (plus jaune) correspond à l'azimut et l'élévation exacts de la source acoustique.

3.5 Conclusion

Dans ce chapitre on a exploré le problème de la localisation des sources acoustiques à partir des représentations géométrique et analytique. On a également évoqué les paramètres des signaux de sources acoustiques devant être déterminés en vue d'un mixage avec le signal d'un microphone principal HOA. Dans le cadre de ce travail on ne dispose que des signaux captés par les microphones d'appoint et le microphone principal.

Nous avons proposé une méthode d'estimation du retard basée sur la méthode classique d'intercorrélation. On a montré que pour l'estimation du retard on pouvait utiliser le signal capté par le microphone d'appoint et la première composante HOA *w* qui peut être considérée comme un microphone de type « omni ». L'application de la fonction d'intercorrélation (54) pour ces deux signaux peut présenter des problèmes, en particulier en présence de plusieurs sources ou de réverbération. Pour éviter ces problèmes un lissage temporel de la fonction d'intercorrélation a été proposé.

La deuxième étape de l'analyse consiste à déterminer la position de la source acoustique en utilisant le retard précédemment estimé. Cette estimation de la position de la source peut être effectuée dans le domaine temporel ou dans le domaine fréquentiel. Dans le domaine temporel on a déduit la forme analytique des équations d'estimation (63) - (65) basée sur l'utilisation des signaux Ambisonic d'ordre 1. On a également proposé l'analyse dans le domaine fréquentiel pour chaque bin fréquentiel (70), (71) dont l'avantage est d'utiliser au lieu du produit scalaire entre des complexes un produit simple entre des réels.

Pour trouver les estimations robustes pendant l'analyse fréquentielle on a proposé une méthode de recherche des fréquences interférées d'une source acoustique par rapport aux autres. Cette méthode basée sur le rapport des amplitudes de chaque source acoustique (chapitre 3.3) permet de dire avec quelles fréquences on a des sources acoustiques qui s'interfèrent et donc de sélectionner les fréquences non-interférées pour l'analyse fréquentielle.

Afin de juger de la qualité de l'estimation d'un des paramètres on a proposé d'utiliser des descripteurs basés sur les caractéristiques des signaux telles que l'énergie ou les pics d'intercorrélation. Différentes combinaisons de descripteurs peuvent composer des indices de confiances qui permettent juger de la robustesse de l'estimation sur une échelle de 0 à 1. Ces indices de confiance sont implémentés dans un algorithme (chapitre 4) et testé avec des scènes synthétique et réelles (chapitre 5).

4 Implémentation de l'algorithme

Dans ce chapitre, on présente l'algorithme d'estimation des paramètres permettant de mixer les signaux captés par un microphone principal HOA avec ceux de microphones d'appoint. Cet algorithme est basé sur les méthodes présentées dans le chapitre 3. Cet algorithme peut être implémenté dans un plugin audio et ainsi permettre à un ingénieur du son de disposer d'une fonction d'assistance automatique au mixage intégrée à une station de travail d'édition audio (DAW : Digital Audio Workstation).

Selon les protocoles d'implantation admis par la station « hôte », on s'intéressera à la déclinaison de la brique audio en différents formats de plugin audio, comme VST, AAX, AudioUnit, le plus répandu étant le VST (pour Visual Studio Technology [27]). En amont d'une implémentation C/C++ pour un tel plugin, une implémentation en Matlab été réalisée en premier lieu, et selon une architecture semblable. Celle-ci a servi à évaluer et améliorer l'algorithme sur des scènes sonores dont la composition est bien contrôlée et avec différents niveaux de complexité (chapitre 5).

Le présent chapitre s'en tient essentiellement au calcul des estimateurs et descripteurs introduits au chapitre précédent pour ce qui est de la description algorithmique « bas-niveau » et à l'agencement global des modules pour la description « haut-niveau ». L'usage concret des descripteurs pour consolider l'analyse et fournir des paramètres robustes et exploitables par l'ingénieur du son, sur la base des expérimentations sera présenté au chapitre 5.

On considère les étapes d'estimation des paramètres présentées sur la Figure 41 avec les signaux des microphones d'appoint et du microphone principal.

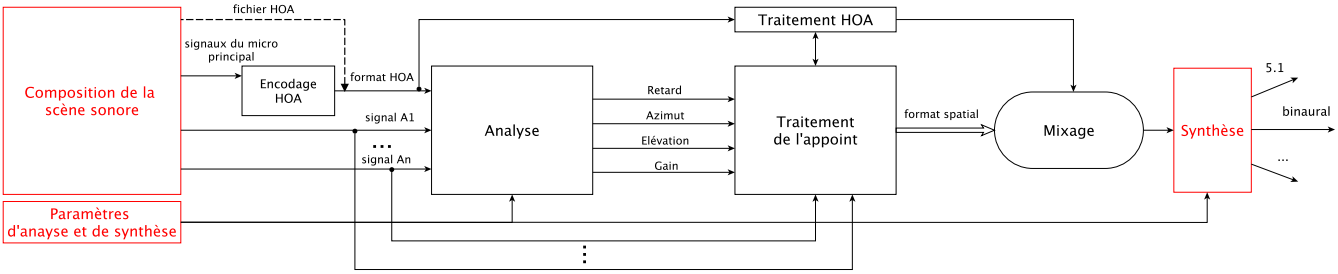

Figure 41. Estimation des paramètres réalisée en Matlab. Par rapport au schéma présenté dans la chapitre 2 (Figure 19) les blocs rouges sont complémentaires dans la réalisation Matlab pour pouvoir créer différentes scènes sonores et écouter le résultat.

Le module de « Composition de la scène sonore » (Figure 41) réalisé en Matlab permet de créer différentes scènes sonores à partir des fichiers audio associés aux microphones d'appoint et au microphone principal. Pendant l'étape de la création de scène sonore simulée le module offre une gestion complexe :

- création de la scène avec une ou plusieurs sources acoustiques avec les microphones d'appoint associés à partir des fichiers audio mono ;
- création de la scène avec plusieurs sources acoustiques à proximité d'un microphone d'appoint associé ;
- intégration du retard, de l'azimut, de l'élévation et du gain simulés pour chaque microphone d'appoint ;
- simulation de l'effet de « diaphonie » pour chaque microphone d'appoint avec la possibilité de contrôler le retard et le gain entre les différentes sources et signaux microphoniques ;
- application d'un effet de réverbération dans le domaine ;
- utilisation du signal au format HOA pour le microphone principal ;
- utilisation du signal multicanal (par exemple, 32 canaux d'enregistrement du microphone Eigenmike) en tant que signal du microphone principal ;
- simulation d'un enregistrement HOA à partir des signaux de microphones d'appoint pour créer le signal du microphone principal ;
- création du signal de microphone principal à partir des signaux des microphones d'appoint et d'une réponse impulsionnelle au format HOA.

Pour effectuer l'analyse et voir les résultats après la synthèse le programme propose de régler les « Paramètres d'analyse et de synthèse » (Figure 41) suivants :

- sélection des microphones d'appoint pour l'analyse ;
- sélection d'un segment temporel dans une séquence audio pour l'analyse ;
- définition d'une distance maximale D_{max} entre le microphone d'appoint et le microphone principal ;
- activation/désactivation de l'analyse fréquentielle en complément de l'analyse dans le domaine temporel ;
- sélection de la taille des trames pour l'analyse, et d'autres paramètres ayant trait à l'analyse par bloc de signal (voir Section 4.1) ;
- sélection des différents paramètres pour l'analyse des pics d'intercorrélation
- définition du descripteur E_a ;
- sélection du format de la restitution sonore (par exemple, binaural) à la sortie du module « Synthèse » (Figure 41).

4.1 Structuration des signaux pour l'estimation

A l'intérieur du bloc « Analyse » l'algorithme effectue l'estimation dans les domaines temporel et/ou fréquentiel en prenant en compte les indices de confiance (chapitre 3.4). A la sortie du bloc « Analyse » on obtient idéalement les paramètres considérés comme bien estimés pour les fournir à l'ingénieur du son afin d'effectuer le mixage.

Pour estimer le retard entre deux signaux, les définitions suivantes sont introduites :

- le « signal de référence » correspond au signal capté par le microphone d'appoint, considéré comme plus net par rapport à celui capté par le microphone principal, car plus proche de la source à mixer. Il contient cependant des éléments susceptibles d'être identifiés également dans le « signal d'observation », en général de façon altérée et/ou mélangée à d'autres signaux ;
- le « signal d'observation » correspond généralement à la composante omnidirectionnelle *w*. Ce signal doit comporter une partie du « signal de référence » décalé temporellement et doit être représentatif de ce qui est capté par le microphone principal (Figure 42).

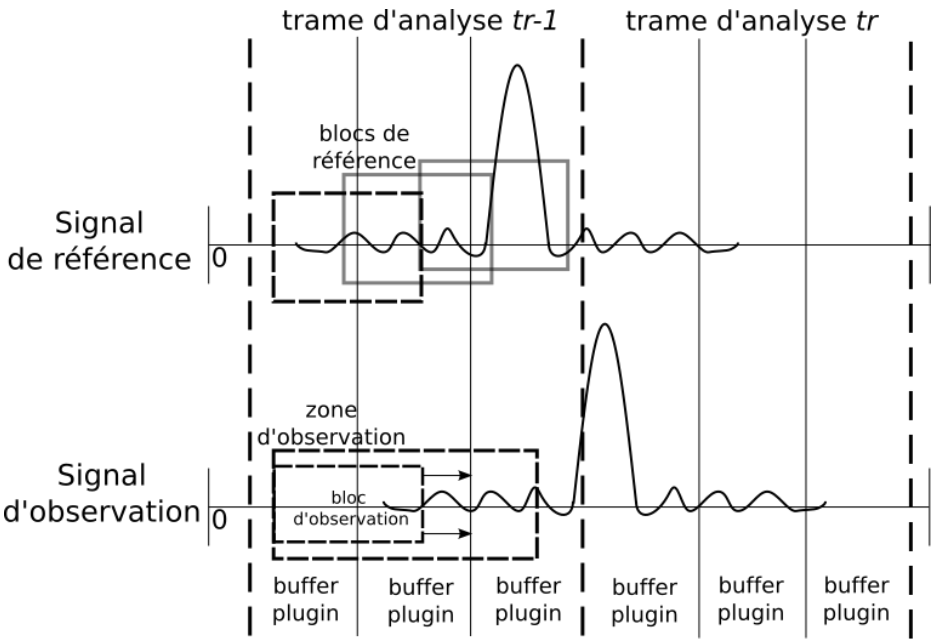

Figure 42. Deux signaux à comparer. Structure et notions.

Le signal de référence et le signal d'observation sont composés des buffers plugin (vecteurs d'échantillons audio), reçues progressivement par la brique d'analyse (Figure 42). La mémorisation de ces buffers plugin permet de constituer une plus large séquence temporelle utilisée pour l'analyse, dénommée « trame d'analyse », dont la taille fixe et suffisamment petite pour refaire régulièrement le traitement. A chaque mise à jour de trame d'analyse par des

buffers plugin, on sélectionne dans le signal de référence un ou plusieurs blocs de référence (Figure 42). Ces blocs peuvent être disjoints, accolés ou se chevaucher. En toute généralité leur taille pourrait être variable de sorte à s'adapter au signal, afin par exemple de s'ajuster à une caractéristique intéressante du signal comme un transitoire qui sera plus facilement identifiable dans la zone d'observation.

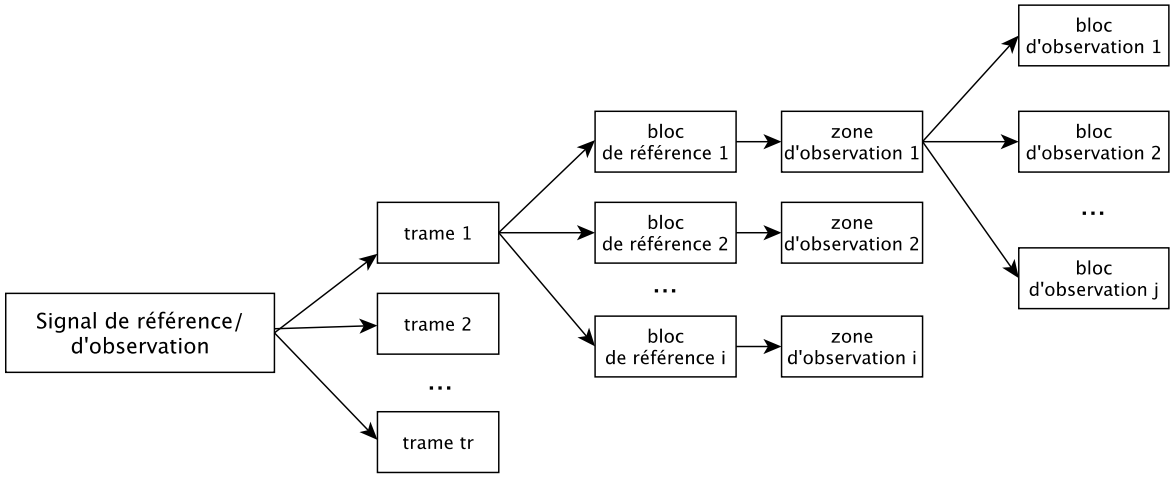

Figure 43. La structure du signal de référence et d'observation représentée par les *tr* trames, *i* blocs de référence, *i* zones d'observations et *j* blocs d'observation.

Au sein d'une trame d'analyse, on passe d'un bloc de référence *i* au bloc suivant *i+1* en se déplaçant d'un certain pas d'échantillons. Dans ce travail, on utilise un pas d'avancement du bloc de référence constant de préférence, et qui peut être soit identique à la taille du buffer (blocs accolés), soit inférieur (blocs se chevauchant comme dans la Figure 42), soit supérieur (blocs disjoints, ce qu'on évitera).

Chaque trame d'analyse (indexée par l'indice *tr*) est constituée d'une ou plusieurs zones d'observation (indexées par l'indice *i*) (Figure 43) relatives aux blocs de référence (chacune d'elles est censée contenir le bloc de référence retardé). La taille de la zone d'observation est donnée par la somme de la taille du bloc de référence et du retard maximal possible τ_{max} entre le microphone d'appoint et le microphone principal qui peut être exprimé à travers la distance maximale possible D_{max} entre les deux microphones : $\tau_{max} = D_{max}/c$ où c est la célérité du son (≈340ms^{-1}). Il faut noter que la taille de la zone d'observation peut être variable en fonction des estimations effectuées (par exemple, si la source n'est que très faiblement mobile, il est inutile d'aller chercher un retard très différent de celui qui a été trouvé précédemment).

Au sein d'une zone d'observation, on définit les blocs d'observation comme des blocs successifs séparés par un certain pas. Ce pas est généralement constant et égal à 1 (cas de

l'intercorrélation classique), mais peut être plus grand (voire variable, voire encore lié à une approche d'optimisation) afin de diminuer la puissance de calcul nécessaire à l'intercorrélation (routine la plus coûteuse de l'algorithme). Les blocs d'observation ne sont introduits que pour expliciter précisément le calcul de similarité. Dans le cas de l'estimation du retard, c'est la fonction d'intercorrélation qui est appliquée pour trouver la similarité entre les blocs et pour l'estimation de la position les équations (63) - (65) sont utilisées.

4.2 Traitement consolidé par bloc

Considérons a_{bRef_i}, le i-ème bloc de référence dans le signal capté par le microphone d'appoint a. Le bloc d'observation de la première composante HOA est désigné par w_{bObs}. Pour calculer la similarité entre les blocs de référence et d'observation, on réécrit la fonction d'intercorrélation normalisée (54) comme suit:

$$\mathrm{C}_i(\tau) = \frac{\langle a_{bRef_i} | w_{bObs} \rangle_{-\tau}}{\|a_{bRef_i}\| . \|w_{bObs}\|_{-\tau}} \tag{97}$$

Le retard entre les deux signaux, pour le i-ème bloc, est donc donnée par :

$$\tilde{\tau}_i = \underset{\tau}{\operatorname{argmax}}\, C_i(\tau) \tag{98}$$

A partir du retard estimé pour le i-ème bloc on peut estimer la position et le gain de la source acoustique :

$$\tilde{\theta}_i = \operatorname{atan2}\left(\langle a_{bRef_i} | y_{bObs} \rangle_{-\tilde{\tau}}, \langle a_{bRef_i} | x_{bObs} \rangle_{-\tilde{\tau}}\right) \tag{99}$$

$$\tilde{\varphi}_i = \arcsin\left(\frac{\langle a_{bRef_i} | z_{bObs} \rangle_{-\tilde{\tau}}}{\eta \cdot \langle a_{bRef_i} | w_{bObs} \rangle_{-\tilde{\tau}}}\right) \tag{100}$$

$$\tilde{g}_i = \frac{\langle a_{bRef_i} | w_{bObs} \rangle_{-\tilde{\tau}}}{\|a_{bRef_i}\|^2} \tag{101}$$

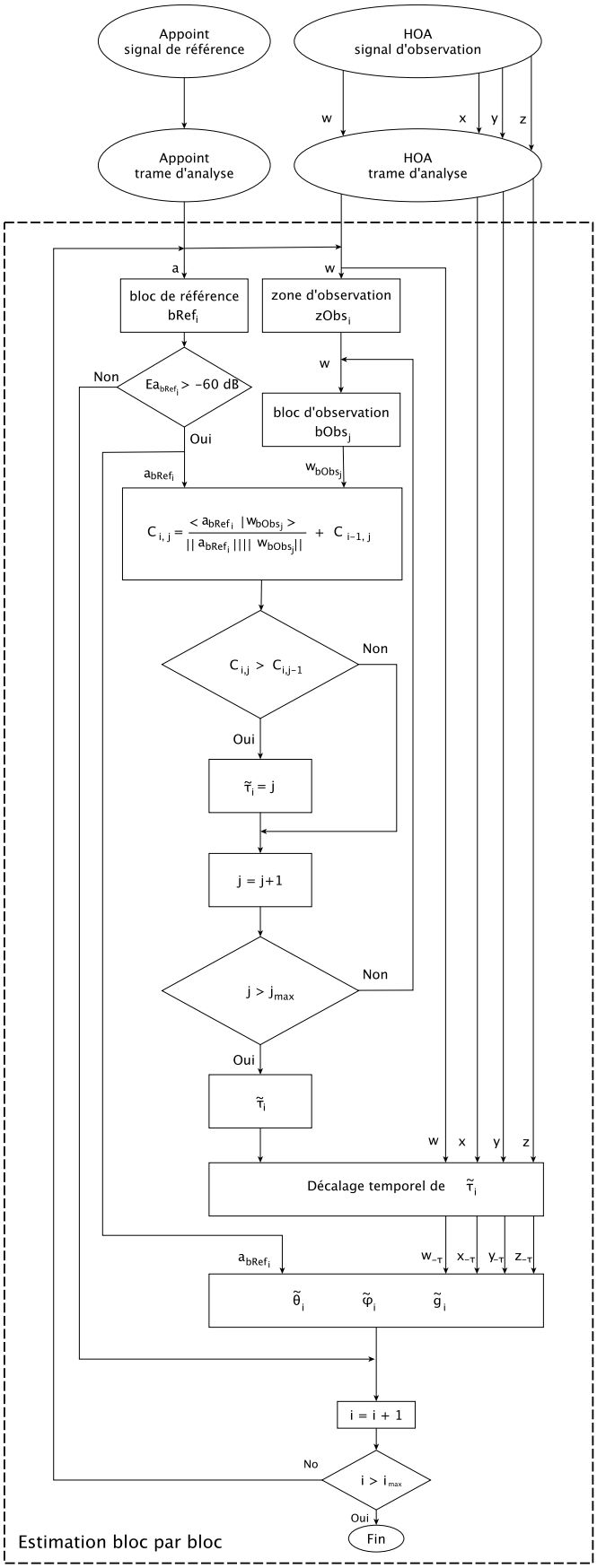

Figure 44. Schéma-bloc de l'estimation des paramètres bloc par bloc dans le domaine temporel pour une trame d'analyse.

La procédure estimation des paramètres par blocs dans le domaine temporel est illustrée sur la Figure 44. L'algorithme utilise un premier descripteur $E_{a_{bRef}}$ correspondant à l'énergie d'un segment de signal du microphone d'appoint analysé. Si ce descripteur est sous un certain seuil (par exemple, -60 dB), on considère que le signal est principalement constitué de bruit et le bloc est ignoré. Par la fonction (98) appliquée au bloc de référence et d'observation le retard $\tilde{\tau}_i$ est ensuite estimé pour le i-ème bloc. Avec le retard $\tilde{\tau}_i$ estimé, les composantes HOA x, y, z sont décalées de $\tilde{\tau}_i$ pour effectuer l'estimation de la position $\tilde{\theta}_i$, $\tilde{\varphi}_i$ et le niveau de gain $\tilde{g}_i$ avant de passer aux blocs suivants.

Dans le domaine fréquentiel, l'estimation du retard est effectuée de la même façon que dans le domaine temporel (Figure 44). En revanche, pour chaque bloc de référence de chaque signal d'appoint, l'algorithme sélectionne une bande de fréquences dans laquelle l'interférence avec les autres signaux d'appoint est minimale (chapitre 3.3). Cette opération a pour but de rendre plus robuste l'estimation de la position. Par analogie avec la notation mathématique utilisée dans ce chapitre pour le domaine temporel, on considère un segment du signal d'appoint $A_{bRef_{i,f_q}}$ pour le i-ème bloc de référence et pour la fréquence f_q « non-interférée ». On considère également le bloc d'observation correspondant pour les différentes composantes HOA à la fréquence f_q : $W_{bObs_{i,f_q}}$, $X_{bObs_{i,f_q}}$, $Y_{bObs_{i,f_q}}$, $Z_{bObs_{i,f_q}}$. En appliquant les équations (70), (71) la position peut être estimée comme suit :

$$\tilde{\theta}_n = \operatorname{atan}\left(\mathfrak{R}\left(\frac{\left(Y_{bObs_{i,f_q}}\right)_{-\tilde{\tau}}}{\left(X_{bObs_{i,f_q}}\right)_{-\tilde{\tau}}}\right)\right) \tag{102}$$

$$\tilde{\varphi}_n = \arcsin\left(\mathfrak{R}\left(\frac{\left(Z_{bObs_{i,f_q}}\right)_{-\tilde{\tau}}}{\eta \cdot \left(W_{bObs_{i,f_q}}\right)_{-\tilde{\tau}}}\right)\right) \tag{103}$$

où $\left(W_{bObs_{i,f_q}}\right)_{-\tilde{\tau}}$, $\left(X_{bObs_{i,f_q}}\right)_{-\tilde{\tau}}$, $\left(Y_{bObs_{i,f_q}}\right)_{-\tilde{\tau}}$, $\left(Z_{bObs_{i,f_q}}\right)_{-\tilde{\tau}}$ sont les blocs d'observations de chaque composante HOA pour une fréquence f_q et décalés du retard $\tilde{\tau}$ estimé.

Le schéma-bloc de l'algorithme d'estimation pour un bloc donné dans le domaine fréquentiel est présenté sur la Figure 45. L'estimation du retard est effectuée de la même façon que pour le domaine temporel (Figure 44). Ensuite les signaux (bloc de référence et blocs d'observation décalés de $\tilde{\tau}_i$) sont transformés dans le domaine fréquentiel par FFT pour passer vers une étape de sélection des fréquences non-interférées. Cette étape de sélection utilise tous les blocs de référence des autres microphones d'appoint. Pour chaque fréquence et chaque

microphone d'appoint, l'amplitude du spectre est comparée avec l'amplitude du spectre des blocs de référence des autres microphones d'appoint. Le rapport en dB entre deux amplitudes de deux blocs de référence associés aux deux microphones d'appoint indique le degré d'interférence des signaux pour la fréquence choisie. La position des sources est ensuite effectuée en appliquant les équations (102) et (103). pour les fréquences choisies.

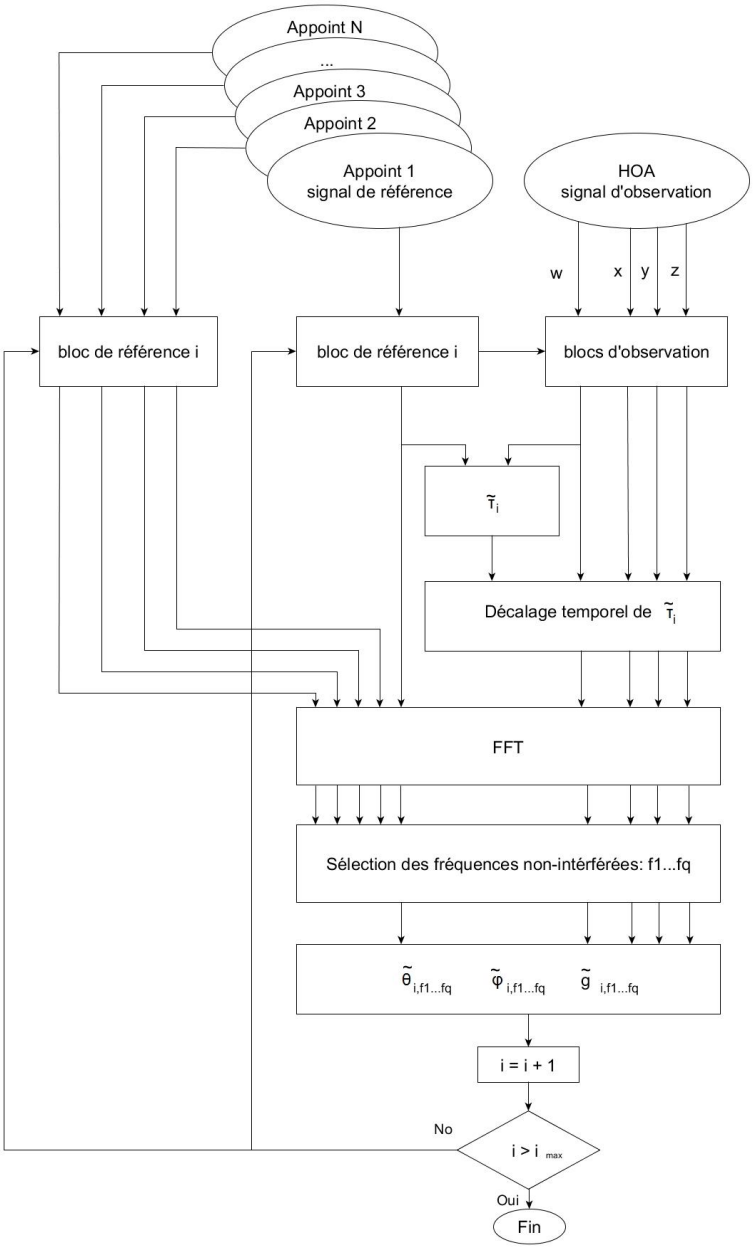

Figure 45. Schéma-bloc de l'estimation des paramètres par blocs dans le domaine fréquentiel.

Le traitement par bloc dans les domaines temporel et fréquentiel possède une structure commune : à partir des données d'entrée, le retard est tout d'abord estimé puis à partir de ce retard l'algorithme estime la position de la source. Dans le domaine temporel, chaque bloc de référence fournit un retard estimé et deux paramètres de position : azimut et élévation. Par

contre, dans le domaine fréquentiel, avec le retard estimé pour chaque bloc, l'algorithme propose une plage d'estimation d'azimut et d'élévation pour les fréquences choisies.

4.3 Traitement consolidé par trame d'analyse

Les schémas-blocs présentés dans la section précédente montrent un premier niveau d'analyse pour un bloc de référence donné, qui ne prend pas en compte les indices de confiance évoqués dans le chapitre 3.4. Le traitement par trame considéré comme le dernier niveau d'analyse afin de fournir les paramètres bien estimés à l'utilisateur est composé de deux étapes : estimation du retard avec les indices de confiance associés et estimation de la position avec les indices associés (Figure 46). A l'entrée les signaux de microphone d'appoint et du microphone principal sont divisés en trames d'analyse (dans le cas du traitement en temps réel l'algorithme reçoit une trame de signal et pas le signal entier). Ensuite la première étape d'analyse est l'estimation du retard par bloc à l'aide de la méthode décrite dans la Section 4.1 (Figure 45). Dans chaque trame d'analyse il est possible d'avoir plusieurs blocs de référence et, par conséquent, plusieurs estimations du retard. Les indices de confiance permettent de sélectionner la meilleure estimation pour une trame donnée.

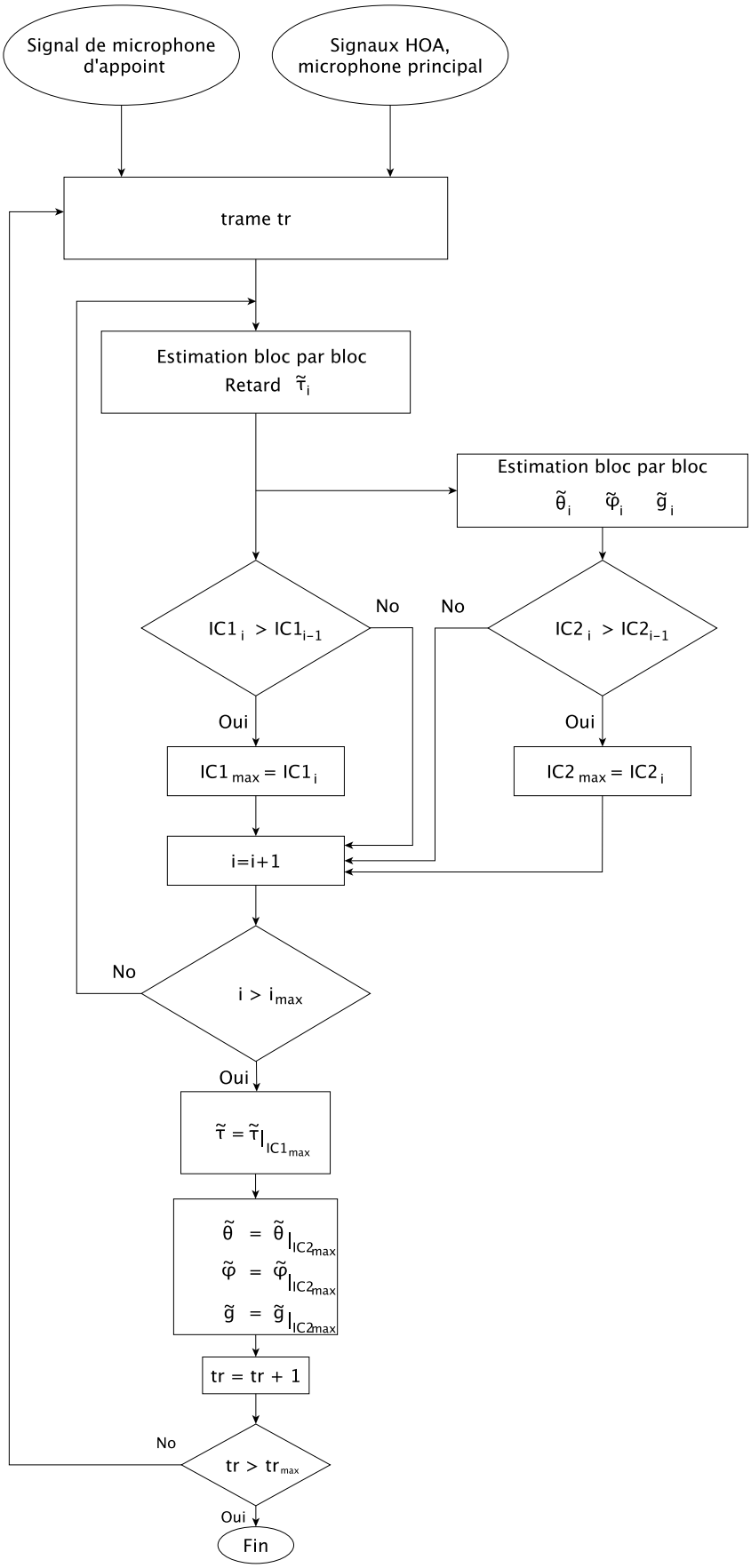

Figure 46. Schéma-bloc de l'algorithme final d'estimation de paramètres par trame d'analyse.

L'estimation du retard est accompagnée par l'indice de confiance $IC1$ qui est le produit de deux descripteurs : l'amplitude maximale de la fonction d'intercorrélation lissée et le ratio de deux premiers pics d'intercorrélation (Section 3.4, équation (81)). D'autre part, une fois le retard estimé, l'algorithme effectue l'estimation de la position avec l'indice de confiance $IC2$ qui est le produit de l'amplitude maximale de la fonction d'intercorrélation et du rapport d'énergie entre le signal du microphone d'appoint et de la première composante HOA w (82).

Une fois l'estimation effectuée pour tous les blocs, l'algorithme sélectionne les meilleures estimations de paramètres en se fondant sur la valeur des indices de confiance (Figure 46). Une seule valeur de retard, d'azimut, d'élévation et de gain sont donc sélectionnés pour une trame donnée.

4.4 Mixage HOA

L'étape finale de la chaîne de production audio pour chaque microphone d'appoint est l'opération de mixage avec les signaux HOA afin d'être restitué vers un système de diffusion du son. Le retard estimé $\tilde{\tau}$ est appliqué au signal du microphone d'appoint pour pouvoir synchroniser ce signal avec celui du microphone principal (Figure 47).

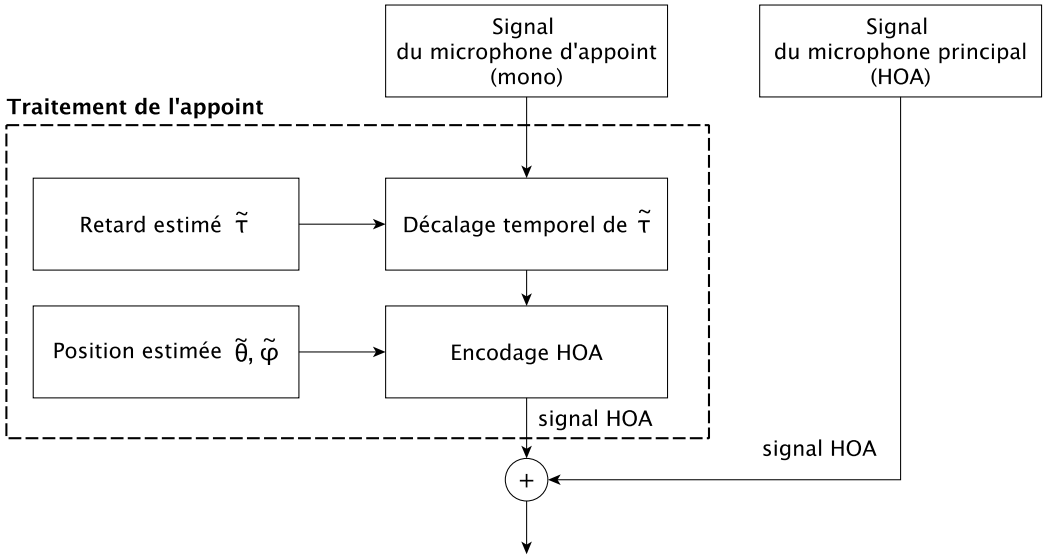

Figure 47. Mixage du signal de microphone d'appoint avec le signal du microphone principal.

Le signal mono du microphone d'appoint décalé de $\tilde{\tau}$ est ensuite encodé au format HOA en prenant en compte la direction définie par l'azimut et l'élévation estimés $(\tilde{\theta}, \tilde{\varphi})$. Avec $\tilde{\theta}, \tilde{\varphi}$ le module d'encodage HOA génère les composantes HOA B_{mn}^{σ} à partir des fonctions harmoniques sphériques $Y_{mn}^{\sigma}(\tilde{\theta}, \tilde{\varphi})$ (voir l'Annexe 2, Extension vers les ordres supérieurs) en respectant l'ordre HOA qui correspond à celui des signaux provenant du microphone principal.

Les signaux HOA encodés à partir du signal du microphone d'appoint sont ensuite mixés aux signaux HOA provenant du microphone principal. Cette étape est effectuée pour chaque microphone d'appoint et permet d'obtenir le mixage final au format HOA pour toute la scène sonore. Les signaux HOA obtenus peuvent alors être diffusés sur casque ou haut-parleurs (voir Annexe 2, Décodage HOA).

4.5 Conclusion

Dans ce chapitre l'algorithme d'estimation des paramètres est présenté sous forme de schémas-blocs. L'analyse est divisée en deux étapes de traitement : par bloc et par trame. Si le terme bloc est introduit pour optimiser et faciliter des calculs pendant l'estimation, la trame est liée au buffer plugin qui fournit les signaux pour l'analyse. Le traitement par bloc consiste à calculer les paramètres à l'aide des méthodes présentées dans les sections 3.2 et 3.3. Pendant cette étape l'algorithme fournit seulement les paramètres estimés par bloc sans avoir la sélection des meilleures estimations. En revanche le traitement par trame, qui inclue le traitement par bloc, effectue aussi les calculs des indices de confiance pour les paramètres estimés. Se basant sur cette information, l'algorithme fournit pour chaque trame un ensemble de paramètres considérés comme bien estimés (un retard, azimut, élévation et gain). Dans le cas de l'analyse d'un signal en temps réel (par exemple pendant l'enregistrement dans un DAW), le traitement par trame est capable de suivre le changement de la scène sonore (par exemple les sources acoustiques qui changent leurs positions au cours du temps) et de fournir, pour chaque trame des paramètres estimés aux ingénieurs du son afin de les aider à ajuster les paramètres de mixage.

5 Test de performance

5.1 Motivation et démarche

Pour résumer, le module d'assistance automatique au mixage HOA avec les signaux de microphones d'appoint fournit les paramètres estimés (retard, azimut et élévation) à l'ingénieur du son afin que ce dernier les ajuste pour effectuer le mixage. Tous les signaux sont divisés en trame dont la taille est associée au buffer d'un host de « plugin » audio. Chaque trame fournit des paramètres estimés au cours de temps. Une fois le retard estimé, la deuxième étape consiste à estimer la position de la source acoustique exprimée par l'azimut et l'élévation.

Nous allons maintenant illustrer l'efficacité de l'algorithme et des méthodes proposées en termes d'estimation du retard et de la position angulaire, trame après trame mais aussi plus globalement, sur la base de l'algorithme décrit au chapitre précédent et implémenté sous Matlab. Du code a été développé en complément, d'une part pour générer les scènes sonores simulées, et d'autre part pour visualiser les résultats d'analyse sous différentes formes, en impliquant quand il le faut des indicateurs *a priori* pertinents. A cet effet, il est choisi plusieurs scènes sonores couvrant diverses conditions acoustiques afin d'apprécier la robustesse de l'algorithme mais aussi les difficultés d'estimation dans les cas plus complexes.

Un des buts de ce chapitre est d'explorer le comportement des différents descripteurs évoqués au chapitre 3, avec l'idée de construire sur cette base des indices de confiance et de fiabiliser l'exploitation des estimations. Voici comment ils sont considérés. Pour exclure de l'analyse des endroits de silence ou de respiration et éviter ainsi de mauvaises estimations, on applique le premier descripteur E_{a_n} en utilisant l'équation (77) avec un seuil arbitraire de -60dB sur l'énergie du signal, en dessous duquel le module d'estimation ne prend pas en compte le signal. Le seuil pourrait être réglé manuellement par l'ingénieur du son en fonction des signaux qu'il observe. Ensuite le rapport entre les pics maximaux $R_{\mathrm{II/I}}$ de la fonction d'intercorrélation lissée et son amplitude $C'(\tilde{\tau})$ sont considérés pour l'estimation du retard. La norme du vecteur vélocité r_v, pressenti comme étant le descripteur le plus prometteur associé à l'estimation angulaire, est présentée pour chaque série de tests. Les descripteurs complémentaires tels que le rapport d'énergie $E_{a/w}$ entre le signal du microphone d'appoint et la première composante du signal HOA du microphone principal, et la détection des fréquences interférées et non-interférées (chapitre 3.4) pour les signaux des microphones d'appoint sont également introduits dans les tests.

Pour l'ensemble des analyses présentées dans ce chapitre, nous appliquons les paramètres techniques suivants. La fréquence d'échantillonnage est de 48kHz. La taille des trames est de 1024 échantillons et le pas d'avancement de 512 (soit la moitié d'une trame). Pour le lissage temporel de l'intercorrélation, le facteur d'oubli vaut 0.9 appliqué trame après trame soit tous les 512 échantillons, ce qui correspond à un temps de convergence à 90% de 10ms [2].

Evoquons succinctement les compositions sonores traitées.

Une première scène sonore *simulée* est composée de plusieurs sources acoustiques (instruments) qui ont été enregistrées séparément dans une chambre anéchoïque et qui font partie d'une composition musicale de Mozart (un air d'opéra de Don Giovanni).

La scène sonore a été composée à partir des fichiers audio associés aux signaux des sources acoustiques positionnées en espace avec les paramètres prédéfinis (azimut et élévation). Les sources acoustiques positionnées dans l'espace sont supposées être ici à la même position que les microphones d'appoint qui leur sont associés un à un. Cela nous permet d'évaluer l'erreur d'estimation par rapport aux paramètres spatiaux connus (retard, azimut et élévation) ainsi que la robustesse de l'algorithme dans des situations de complexité croissante (sans ou avec réverbération, avec les fréquences interférées d'autres sources acoustiques).

L'évaluation algorithmique est réalisée ensuite sur un deuxième type de contenu, issu cette fois d'un enregistrement de scène sonore réelle. Cela nous met évidemment dans la situation visée, et ajoute comme difficulté des effets acoustiques réels (filtrage, effet de salle plus complexe, diaphonie potentielle entre sources et microphones d'appoint) ainsi qu'un encodage microphonique qui n'est pas idéal.

5.2 Scène sonore simulée

Les simulations portant sur une seule source encodée en milieu anéchoïque constituent un cas trivial et induisent comme on peut l'attendre une estimation parfaite, en termes de retard comme de position angulaire. C'est pourquoi nous présentons tout de suite une situation un peu moins triviale quoique de complexité modérée : celle de deux sources en milieu anéchoïque.

5.2.1 Deux sources acoustiques

Nous analysons ici une scène sonore simulée de 6 secondes, composée de deux sources acoustiques : une voix de soprano lyrique et un basson (à une distance simulée respective d'environ 8m et 10m par rapport au microphone principal cf. Tableau 1). La présence de deux sources acoustiques dans le signal du microphone principal dégrade l'intercorrélation avec chaque signal d'appoint, et accroît les probabilités d'avoir des erreurs d'estimation du retard. Ainsi l'estimation de la position est également perturbée par la source acoustique concurrente

et peut donner de mauvais résultats dans le cas où ces sources acoustiques partagent des composantes fréquentielles à un instant donné.

Sources	Retard (éch.)	Azimut (°)	Elévation (°)	Gain
Soprano	1127	-30	0	1
Basson	1411	45	30	1

Tableau 1. Configuration de la scène sonore simulée. Deux sources acoustiques.

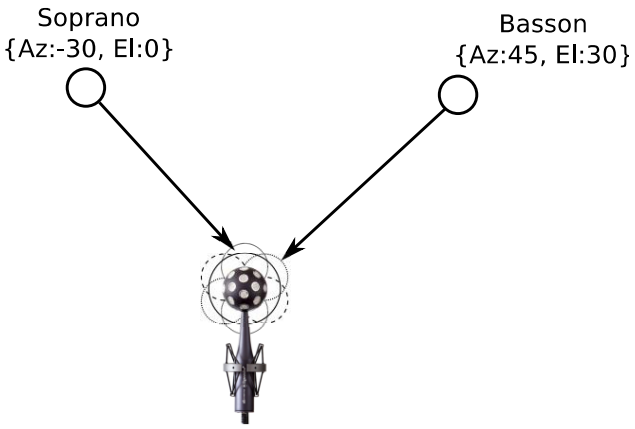

Figure 48. Scène sonore simulée. Deux sources acoustiques.

La scène sonore composée des deux sources acoustiques est représentée par le spectrogramme de la première composante HOA *w*, ainsi que les enveloppes d'énergie de chacune des sources (Figure 49).

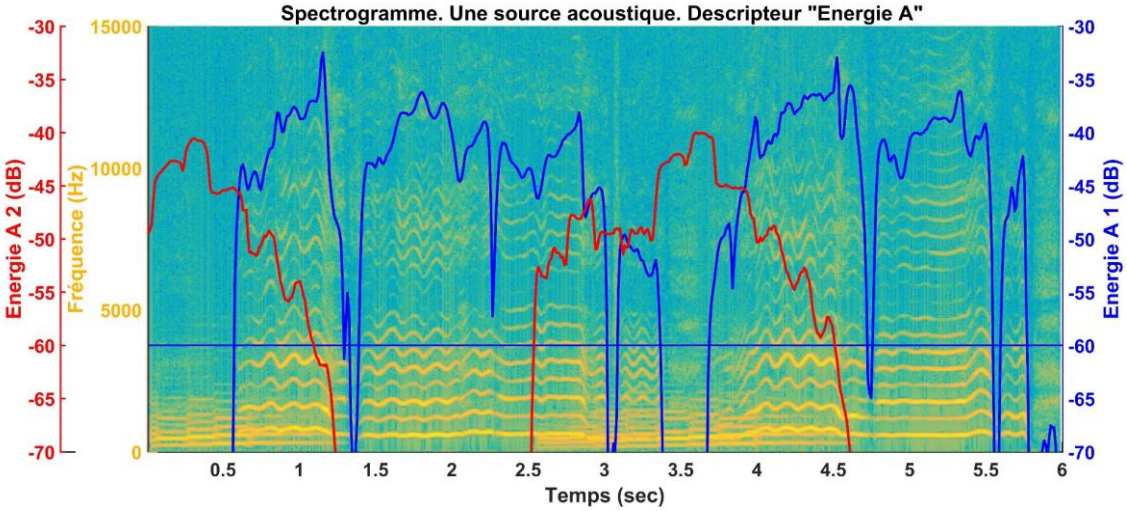

Figure 49. Scène sonore composée de deux sources acoustiques (soprano et basson). Haut : spectrogramme et signal temporelle de la composante HOA *w* ; Bas : Spectrogramme de *w*, l'énergie de chaque source acoustique associée aux microphones d'appoint « A1 » et « A2 ».

Dans la scène sonore présentée, la deuxième source acoustique (Figure 49, enveloppe d'énergie en rouge) perturbe la première source acoustique (enveloppe d'énergie en bleu) sur

plusieurs périodes. L'estimation du retard pour la première source acoustique (Figure 50) basée sur l'intercorrélation lissée montre une convergence correcte vers le retard prédéfini pour la source « soprano » (Tableau 1).

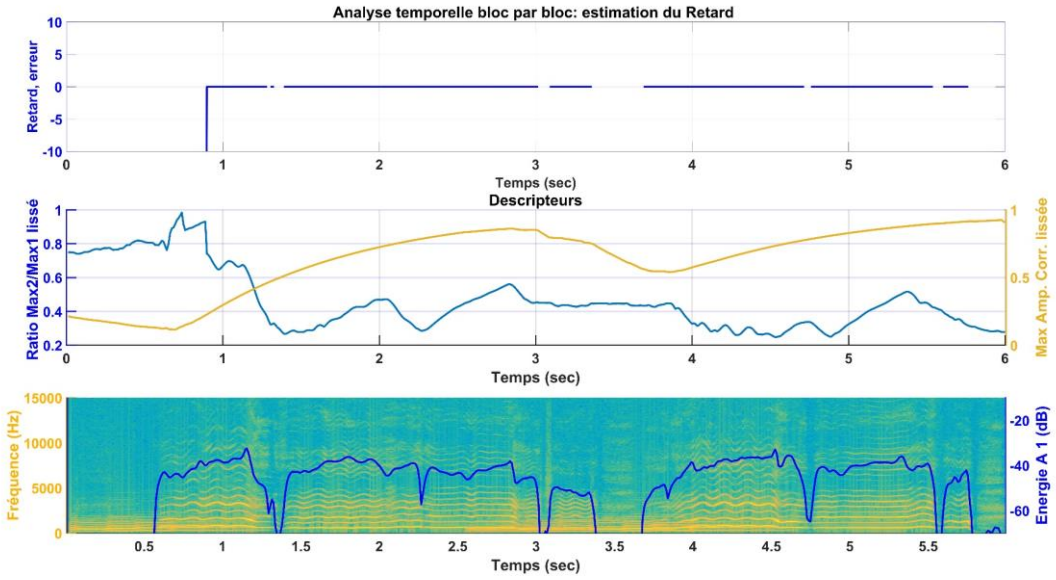

Figure 50. Estimation du retard pour la première source acoustique ('soprano') dans la scène sonore composée de deux sources acoustiques.

La Figure 50 (haut) indique une erreur d'estimation du retard au début de morceau. Du fait que le deuxième instrument (basson) commence à jouer en même temps que le premier (soprano) et que la fonction d'intercorrélation lissée n'a pas encore « accumulé » suffisamment de valeurs calculées auparavant pour faire apparaître un ratio élevé entre les pics maximaux, le module d'estimation donne des erreurs. La contribution de la première source acoustique dans le signal principal est plus élevée que celle de la deuxième source et, par conséquent, on distingue le signal de la première source dans le signal du microphone principal en utilisant la fonction d'intercorrélation. L'amplitude de la fonction d'intercorrélation lissée dans ce cas montre bien qu'au début on a des résultats moins fiables quand une source est perturbée par une autre. Dans les endroits d'absence du signal analysé (par exemple entre 3s et 4s), l'amplitude ne chute pas brutalement, mais progressivement.

De façon similaire à la première source acoustique « soprano », l'estimation du retard pour la deuxième « basson » est présentée sur la Figure 51. Elle est toujours bonne grâce à la fonction d'intercorrélation lissée. Il faut remarquer que l'estimation n'est pas perturbée au début de morceau comme dans le cas de la première source acoustique car le « basson » commence à jouer en premier et en l'absence d'une autre source acoustique on obtient directement une bonne estimation. L'amplitude de la fonction d'intercorrélation lissée dans ce cas est maximale dès le

début avec des résultats d'estimations fiables mais qui décroissent dans les périodes d'absence du signal de la deuxième source acoustique.

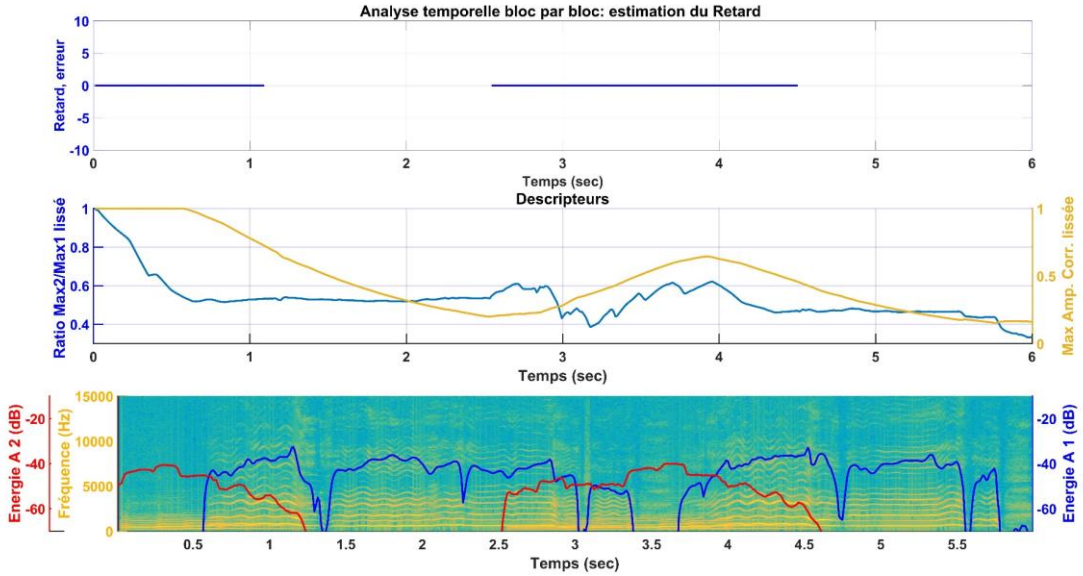

Figure 51. Estimation du retard pour la deuxième source acoustique (« basson ») dans la scène sonore composée de deux sources acoustiques.

On peut constater que l'estimation de la position de chaque source acoustique est fortement perturbée dans les endroits où deux sources acoustiques jouent simultanément. Par exemple, l'azimut et l'élévation de la première source acoustique sont légèrement perturbés au début du morceau (Figure 52, haut) tandis qu'entre la 3ème et la 4ème seconde la contribution de la deuxième source acoustique a un impact sur l'estimation de la position et on obtient des erreurs. Le descripteur $\boldsymbol{r_v}$ (norme du vecteur vélocité) dans cet endroit n'est plus stable, ça veut dire qu'à cet instant le signal de la première source acoustique est fortement perturbé par une autre source acoustique.

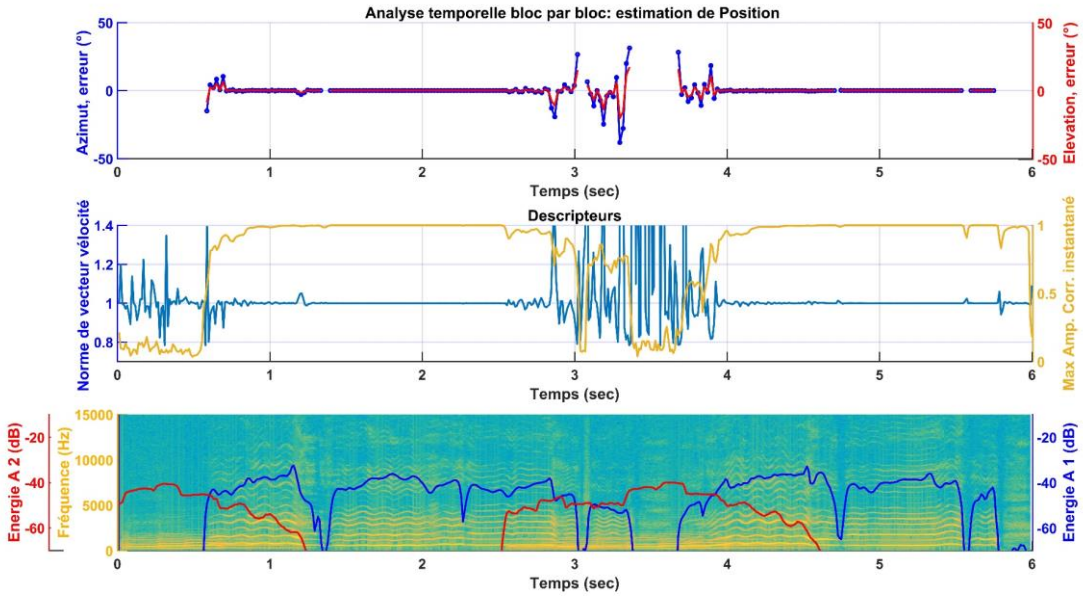

Figure 52. Estimation de l'azimut et d'élévation de la première source acoustique ('soprano') dans la scène sonore composée de deux sources acoustiques.

L'estimation de position de la deuxième source acoustique évoque le même problème avec des erreurs dans les périodes où deux instruments jouent simultanément. En revanche, l'estimation est bonne lorsque le descripteur $\boldsymbol{r_v}$ a des valeurs proches de 1 (Figure 53).

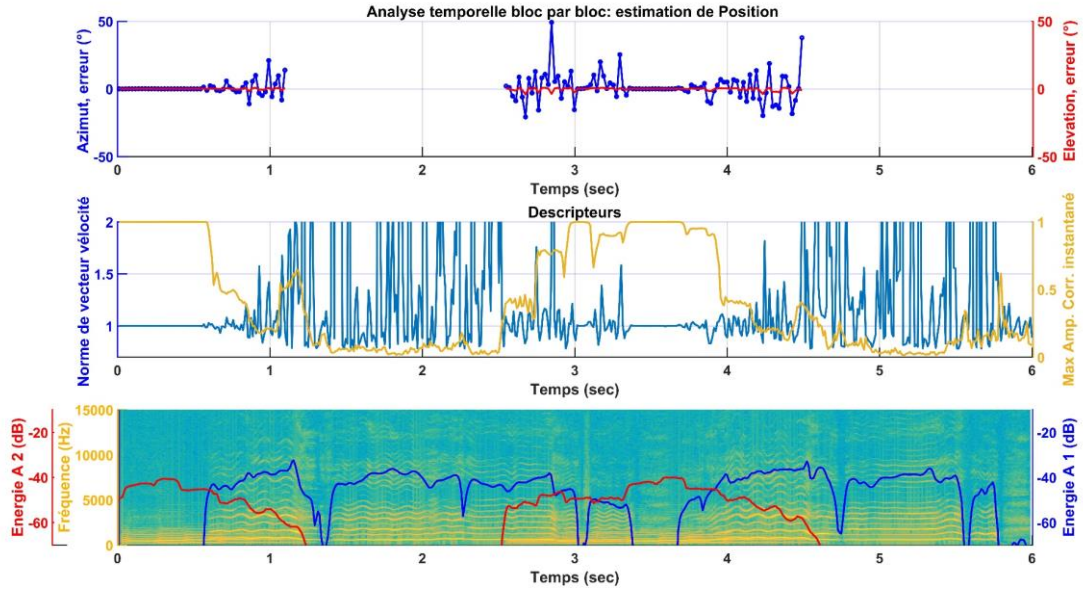

Figure 53. Estimation de l'azimut et d'élévation de la deuxième source acoustique (« basson ») dans la scène sonore composée de deux sources acoustique.

Connaissant les angles d'azimut et d'élévation qui ont servi à la simulation, on peut également calculer et présenter en un seul scalaire l'erreur angulaire en tant qu'écart entre le vecteur vélocité calculé et le vecteur vélocité unitaire de la cible (Figure 54). En prenant en

compte les deux descripteurs $\boldsymbol{r_v}$ et l'amplitude de la fonction d'intercorrélation instantanée (« MaxCorr instantané» sur la Figure 54) on peut constater que pour la première source acoustique ('voix'), la plupart des points correspondant à de bonnes estimations de la position (erreur angulaire minimale) sont concentrés autour de $\boldsymbol{r_v} = 1$ et d'un « MaxCorr instantané » entre 0.8 à 1. De petites variations d'estimations pour $0.9 < \boldsymbol{r_v} < 1.1$ montrent que le vecteur vélocité reste extrêmement proche du vecteur vélocité unitaire de la cible même avec la présence d'une autre source acoustique ce qui justifie que la première source acoustique dans ce cas est dominante.

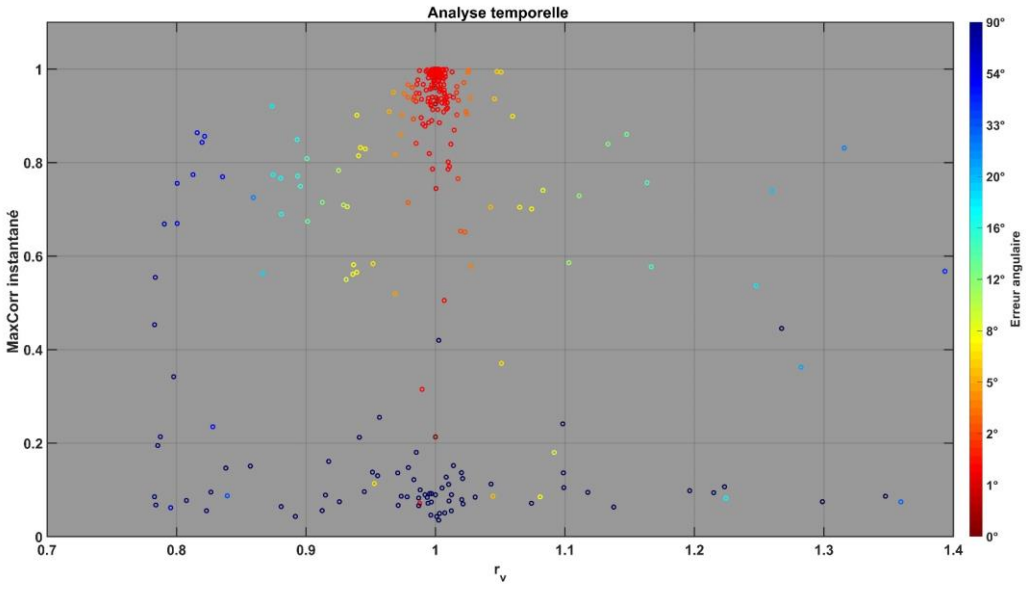

Figure 54. Dans le cas de deux sources en milieu anéchoïque, estimation dans le domaine temporel de la position pour la première source acoustique ('soprano') avec une erreur angulaire et deux descripteurs : norme de vecteur vélocité « r_v » et amplitude maximal de la fonction d'intercorrélation instantanée « MaxCorr instantané». Noter que l'échelle de couleur n'est pas graduée linéairement, afin de mieux distinguer les erreurs faibles et modérées.

Pour la deuxième source acoustique perturbée on peut remarquer que la contribution du descripteur « MaxCorr instantané » dans l'estimation est moindre (Figure 55) par rapport à celle de la première source. De bonnes estimations avec une erreur angulaire sont toujours autour de $\boldsymbol{r_v} = 1$ tandis que les autres sont plus dispersées avec une erreur moyenne de 10° pour les estimations lorsque $0.9 < \boldsymbol{r_v} < 1.1$. On peut donc supposer que la combinaison de deux descripteurs $\boldsymbol{r_v}$ et « MaxCorr instantané » peut fournir un indice de fiabilité de l'estimation.

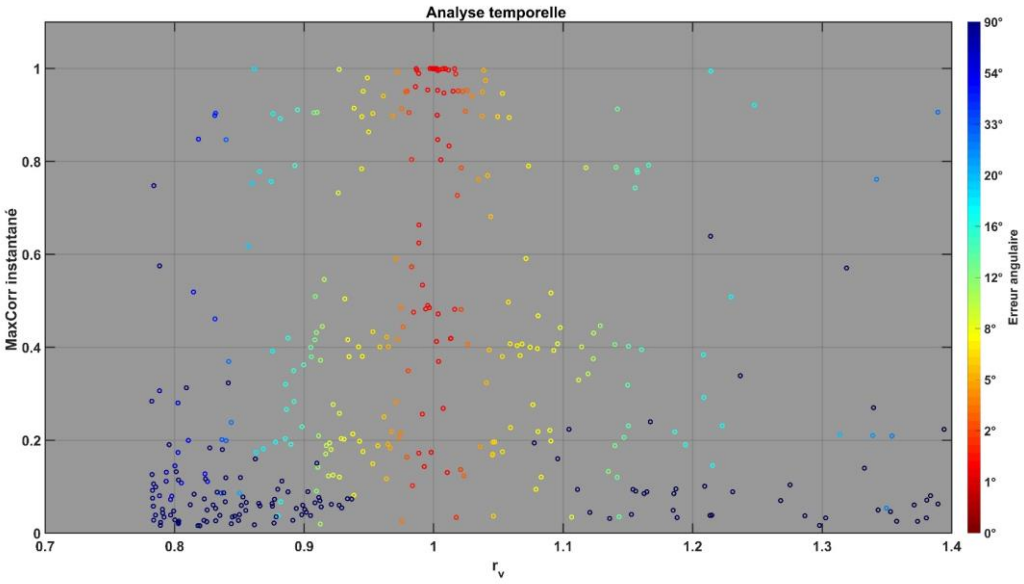

Figure 55. Semblable à la Figure 54 pour la deuxième source acoustique ('basson').

Pour améliorer l'estimation de position on procède à une analyse dans le domaine fréquentiel. L'amélioration est attendue grâce à deux facteurs : d'abord la possibilité de différencier les « bonnes » et « mauvaises » estimations en fonction des fréquences d'après des indicateurs comme la norme du vecteur vélocité ; ensuite, la prise en compte du spectre des sources concurrentes, dans l'hypothèse où celles-ci sont captées par d'autres microphones d'appoint. En présence d'une source concurrente, captée par un autre microphone d'appoint (et sans diaphonie), on fait une séparation des fréquences interférées et non-interférées en regardant le rapport des amplitudes du signal de la source acoustique analysée et l'autre source. On se donne ici le critère suivant : on estime qu'un bin fréquentiel est non-interféré si le rapport d'amplitude est supérieur à 20dB, c'est-à-dire quand le signal de la source concurrente est au moins 10 fois plus faible que le signal de la source analysée. Ce seuil est choisi relativement arbitrairement, mais son impact est mesurable.

Pour la scène sonore analysée, on regarde la représentation temps-fréquence de l'erreur angulaire pour les fréquences interférées (Figure 56, « croix ») et non-interférées (Figure 56, « points ») de la première source acoustique en indiquant les points d'estimation. On peut constater que pour cette scène sonore la séparation en fréquences interférées et non-interférées est plutôt justifiée pour minimiser l'erreur angulaire, c'est-à-dire que pour les fréquences non-interférées les estimations sont fiables : la plupart de temps l'erreur ne dépasse pas 1-2° sauf aux endroits de l'impact d'une autre source acoustique. Aux fréquences interférées on observe au contraire de mauvaises estimations, avec une erreur de plus de 10°.

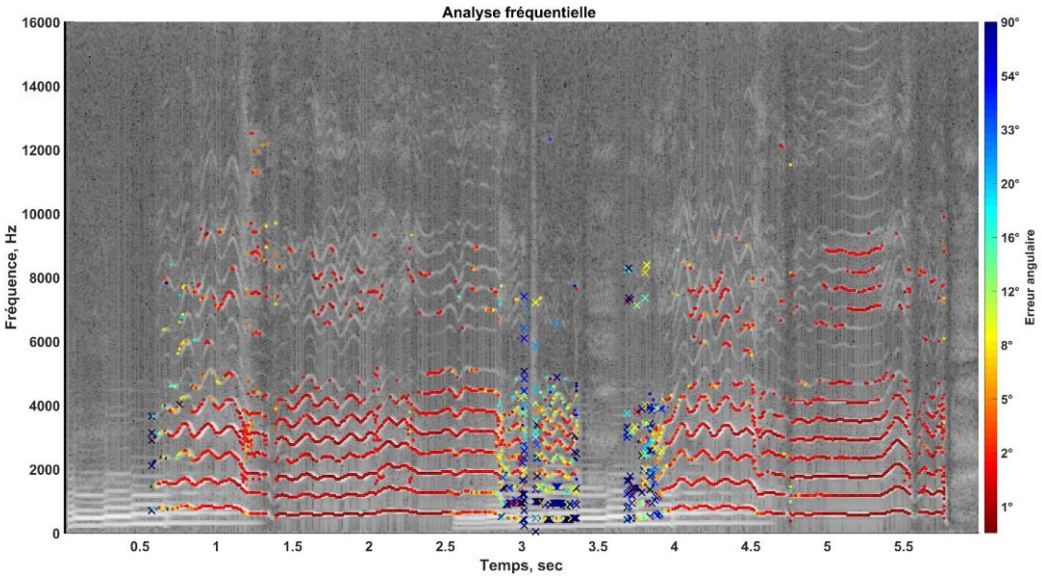

Figure 56. Représentation temps-fréquence de l'erreur angulaire, superposée au spectrogramme de l'extrait analysé de la première source acoustique ('soprano') pour chaque bin fréquentiel interféré (« croix ») non-interféré (« points »).

De même que pour le domaine temporel, on s'intéresse à la norme du vecteur vélocité dans le domaine fréquentiel. Mais dans ce domaine, le vecteur vélocité est en général complexe. On peut regarder séparément la norme de la partie réelle et imaginaire du vecteur vélocité. On rappelle que pour une onde plane seule, la partie réelle est de norme 1 et pointe vers la direction de provenance, quand la partie imaginaire devrait être nulle. Par exemple, pour la première source acoustique dans le cas des fréquences non-interférées on constate qu'il y a une concentration de bonnes estimations avec une erreur angulaire minimale lorsque la norme de la partie réelle de vecteur vélocité est proche de 1 et celle de la partie imaginaire est proche de 0.

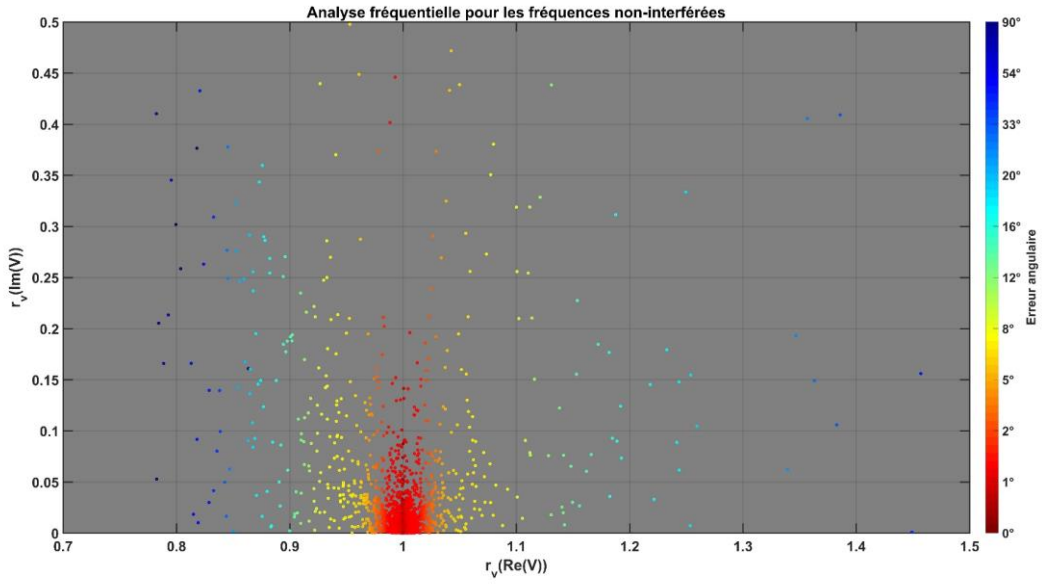

Figure 57. Estimation dans le domaine fréquentiel pour les fréquences non-interférées de la position pour la première source acoustique ('soprano') avec une erreur angulaire et deux descripteurs : norme de la partie réelle et imaginaire du vecteur vélocité.

La deuxième source acoustique qui est plus perturbée (≈3-4s) donne de bonnes estimations pour les fréquences non-interférées mais avec une erreur angulaire un peu élevée par rapport à celle de la première source (Figure 58).

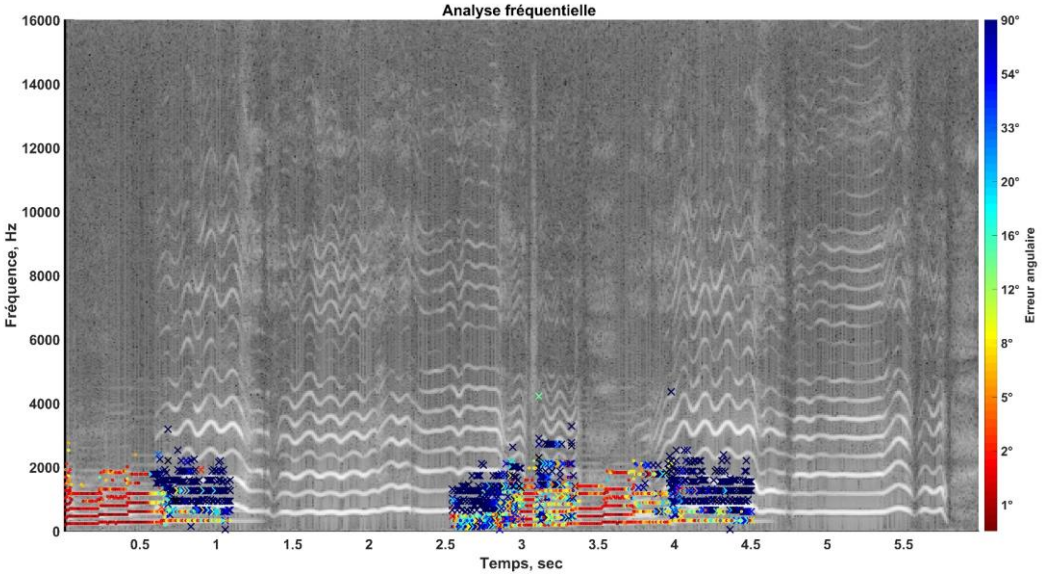

Figure 58. Représentation temps-fréquence de l'erreur angulaire, superposée au spectrogramme de l'extrait analysé de la deuxième source acoustique ('basson') pour chaque bin fréquentiel interféré (« croix ») et non-interféré (« points »).

Les normes des partie réelle et imaginaire du vecteur vélocité dans le cas de la deuxième source acoustique (Figure 59) sont liées à l'erreur angulaire de la même façon que pour la première source. Mais on peut remarquer que même lorsque la norme de la partie imaginaire

commence à croître, l'estimation angulaire reste bonne dès lors que la norme de la partie réelle reste proche de 1.

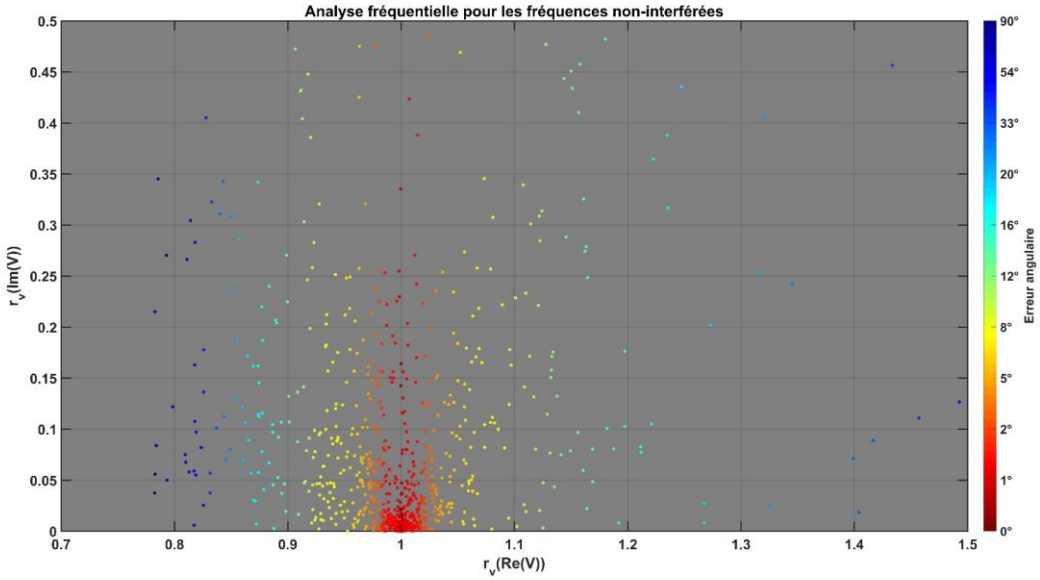

Figure 59. Estimation dans le domaine fréquentiel pour les fréquences non-interférées de la position pour la deuxième source acoustique ('basson') avec une erreur angulaire et deux descripteurs : norme de la partie réelle et imaginaire de vecteur vélocité.

Finalement, dans ce cas l'utilisation de la norme du vecteur vélocité $\boldsymbol{r_v}$ dans les domaines temporel et fréquentiel peut fournir un indice de fiabilité de l'estimation avec la combinaison des différents descripteurs. La séparation des fréquences interférées et non-interférées améliore le résultat en excluant de l'analyse les fréquences partagées par les deux sources acoustiques qui donnent souvent l'estimation erronée.

5.2.2 Une source acoustique avec effet de salle

La réverbération ajoutée à une source acoustique (Figure 60) peut provoquer des erreurs d'estimation à cause de l'influence des ondes secondaires que sont les réflexions acoustiques, sur le signal initial. Comme nous allons le constater, une différence notable, et une difficulté, par rapport aux mélanges de différentes sources sonores est qu'ici les signaux interférents sont très corrélés au son direct que l'on veut localiser. Pour simuler un signal réel on utilise des SRIR (« Spatial Room Impulses Responses ») au format HOA d'une salle de téléconférence (salle Varèse Lannion, Figure 61) et celle de la salle Pleyel (Figure 62) que l'on applique directement au signal monophonique source pour produire les signaux HOA de la scène sonore simulée. Le premier pic de SRIR correspond au front d'onde direct. La localisation de la source acoustique en espace est de -30° en azimut et 5° en élévation pour Varèse, et de 2° en azimut et -9° en élévation pour Pleyel.

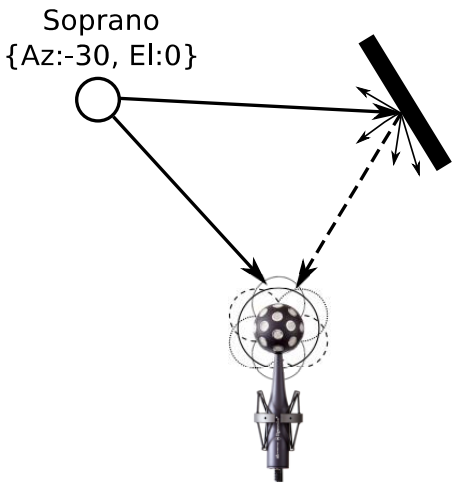

Figure 60. Scène sonore simulée. Une source acoustique avec l'effet de salle.

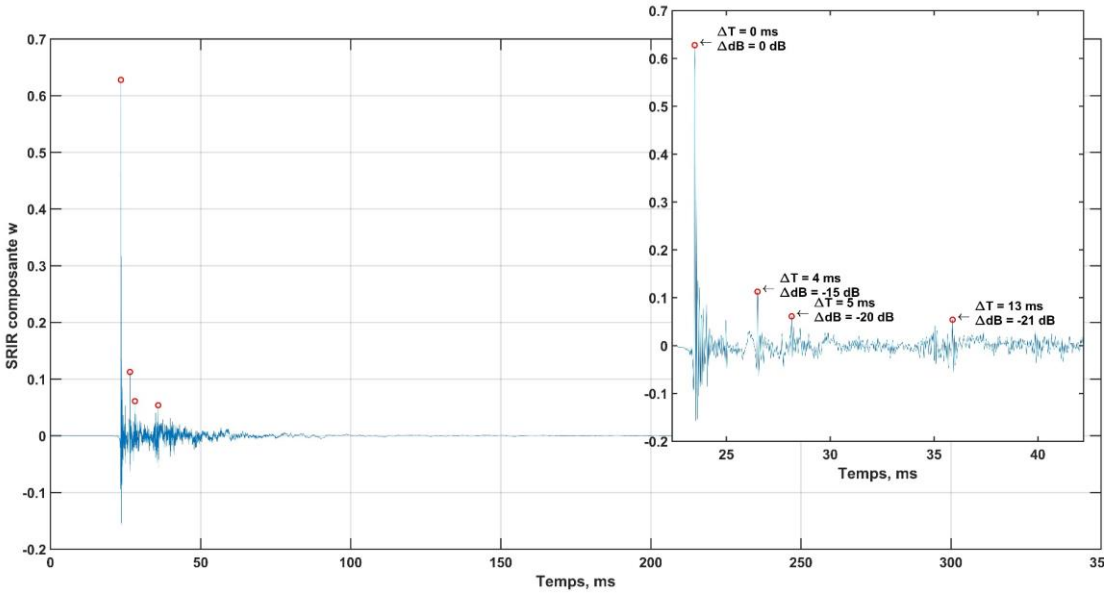

Figure 61. SRIR (première composante HOA) de la salle Varèse à Lannion.

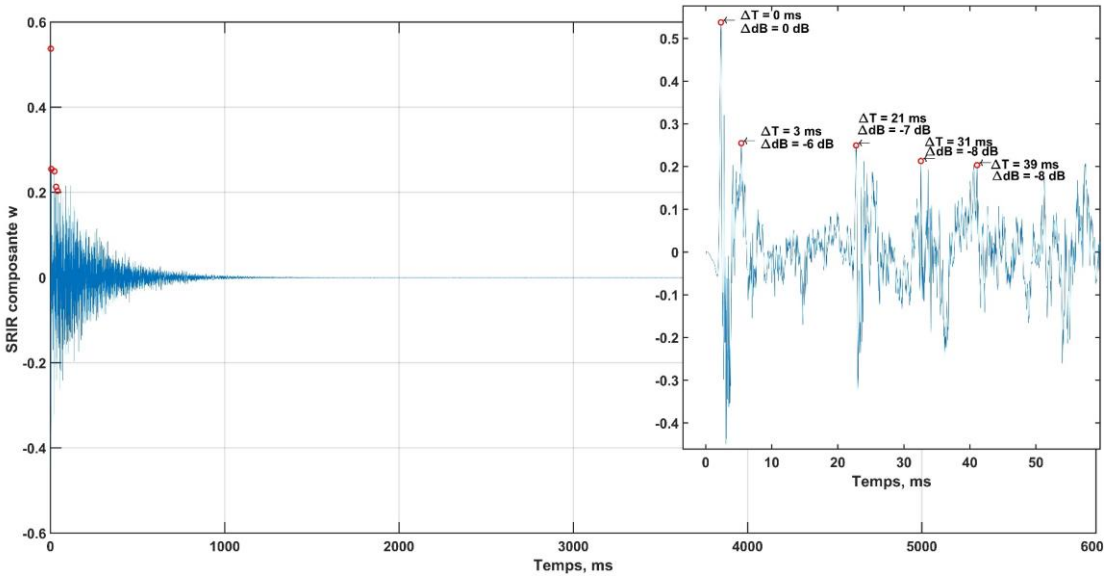

Figure 62. SRIR (première composante HOA) de la salle Pleyel.

L'estimation du retard pour cette scène sonore est toujours bonne (Figure 63) pour la SRIR de la salle Varèse. Par contre, avec la SRIR de la salle Pleyel (Figure 64) les ondes secondaires provoquent une augmentation du niveau des pics secondaires de la fonction d'intercorrélation (Figure 64 , « Ratio Max2/Max1 lissé »). Donc la détection du pic principal de la fonction d'intercorrélation lissée peut être erronée (Figure 64, « Max Amp. Corr. lissée »).

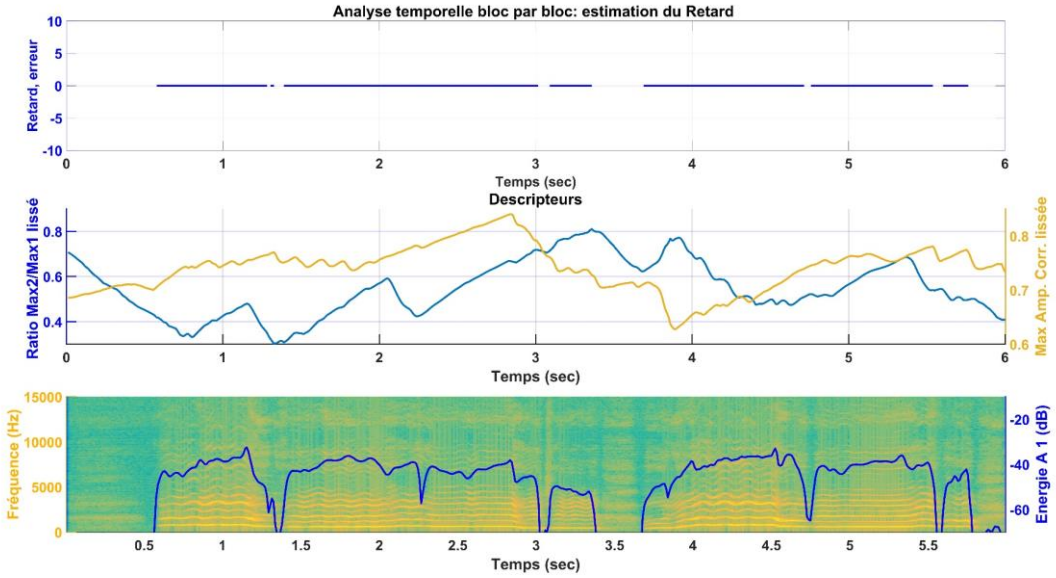

Figure 63. Estimation du retard avec descripteurs d'une scène sonore composée par une source acoustique avec la réverbération de la SRIR de la salle Varèse.

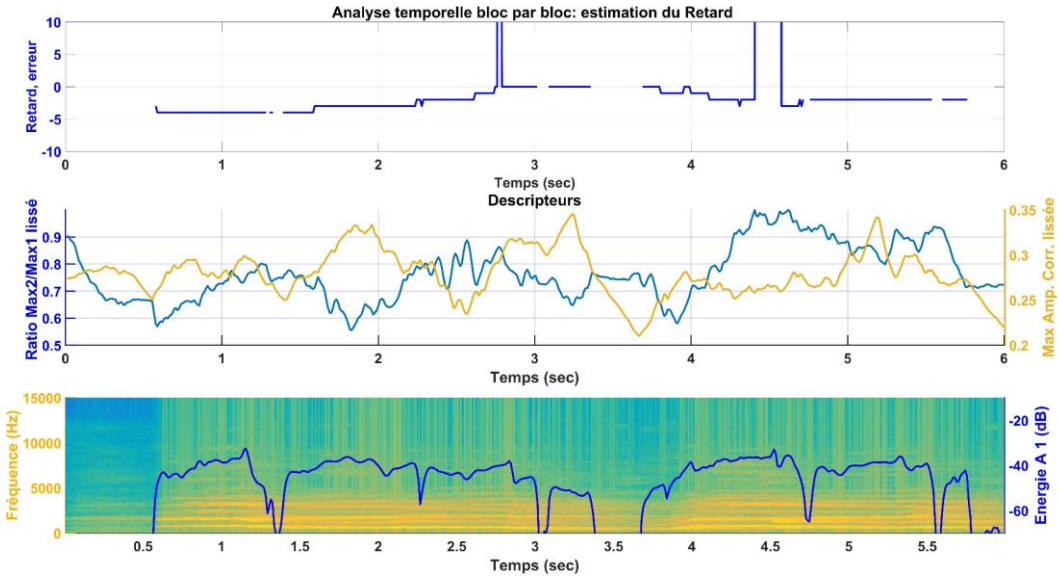

Figure 64. Semblable à la Figure 63 pour la salle Pleyel.

L'estimation de la position est perturbée dans le cas de la salle Varèse (Figure 65) et de la salle Pleyel (Figure 66). Dans le cas de la salle Varèse, l'estimation de l'azimut est moins erronée que celle de l'élévation. Cela s'explique probablement par les réflexions par le sol et par le plafond qui sont parmi les plus précoces, et qui sont de même azimut que le son direct mais d'élévations différentes. Dans les endroits où l'amplitude d'intercorrélation lissée et le descripteur $\boldsymbol{r_v}$ sont proches de 1, on a une estimation avec une erreur minimale (par exemple 2-3 secondes sur la Figure 65).

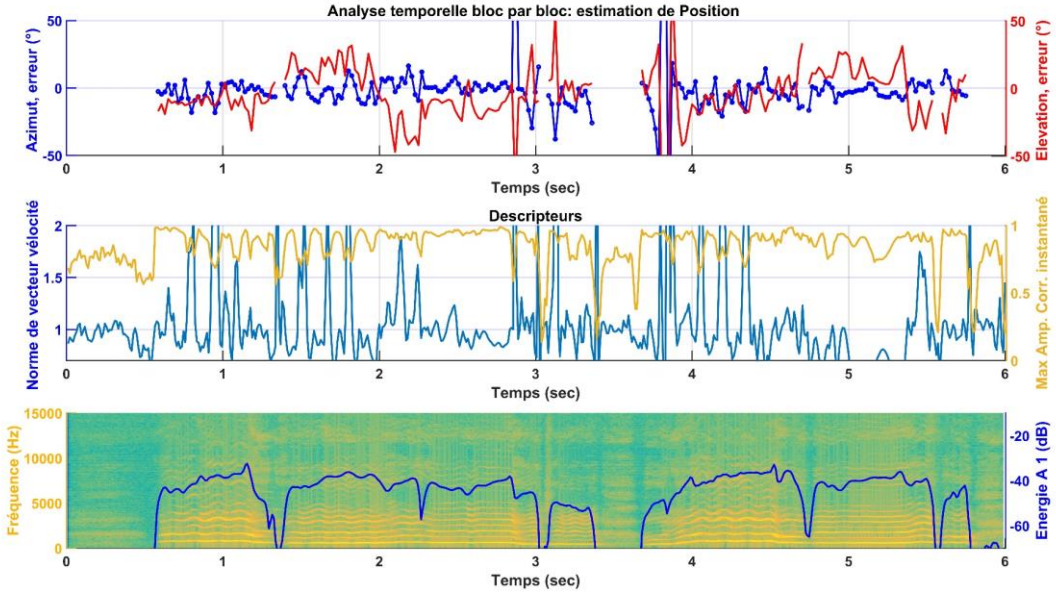

Figure 65. Estimation de la position dans le domaine temporel d'une source acoustique dans la scène sonore avec la réverbération pour la SRIR de la salle Varèse.

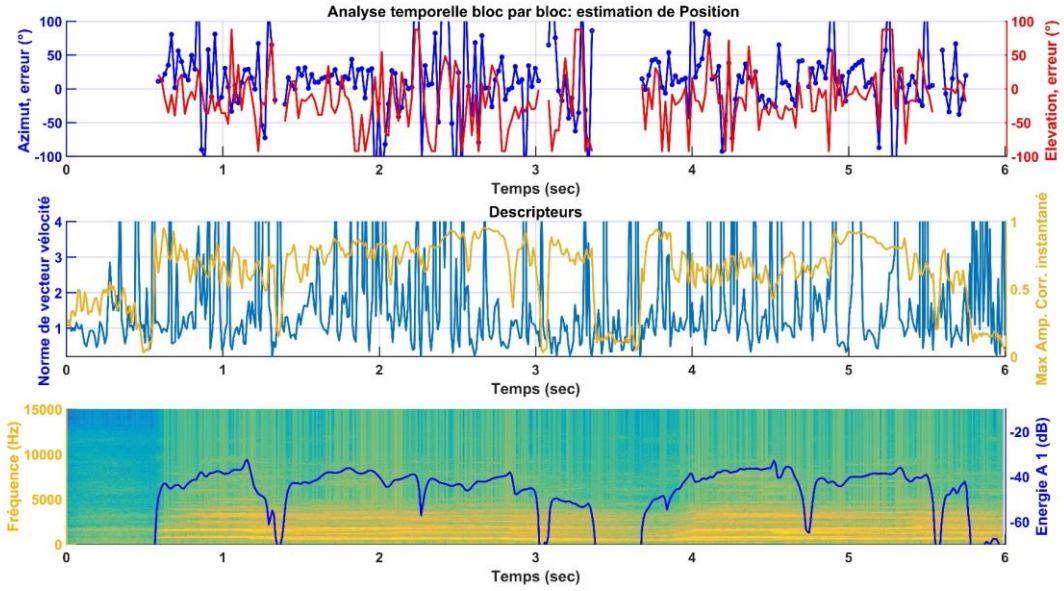

Figure 66. Semblable à la Figure 65 pour la salle Pleyel.

La SRIR de la salle Pleyel pose beaucoup de problèmes pour l'estimation de la position et on peut constater que dans ce cas l'algorithme n'est pas robuste. Il faut préciser que pour la mesure utilisée, le point de captation était à 19 mètres de la source, situation où le son direct se détache relativement peu du son réverbéré (plusieurs réflexions précoces sont à seulement 6 à 8 dB en dessous du son direct comme le montre la Figure 62).

Finalement, l'algorithme estime bien le retard pour le premier cas avec la SRIR de la salle Varèse, et avec de petites perturbations pour la SRIR de la salle Pleyel. Mais l'estimation de la position avec la norme du vecteur vélocité est erronée.

Par analogie avec le cas précédent, on regarde la norme du vecteur vélocité $\boldsymbol{r_v}$ avec la combinaison des autres descripteurs afin de trouver l'estimation fiable pour les deux SRIR.

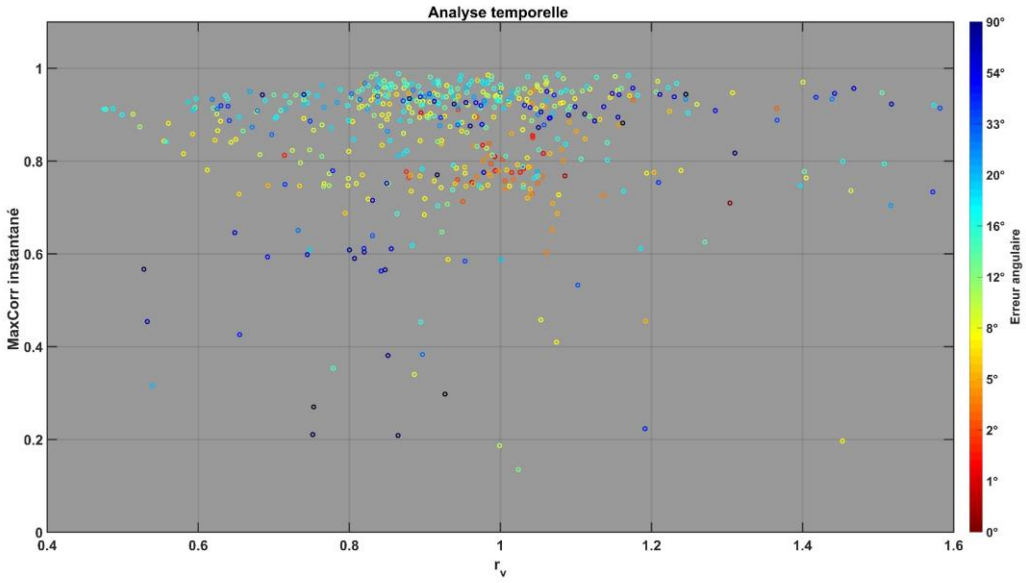

Figure 67. Estimation dans le domaine temporel de la position pour la première source acoustique ('soprano') avec réverbération (SRIR de la salle Varèse) représentée par une erreur angulaire et deux descripteurs : norme du vecteur vélocité « r_v » et amplitude maximale de la fonction d'intercorrélation lissée « MaxCorr ».

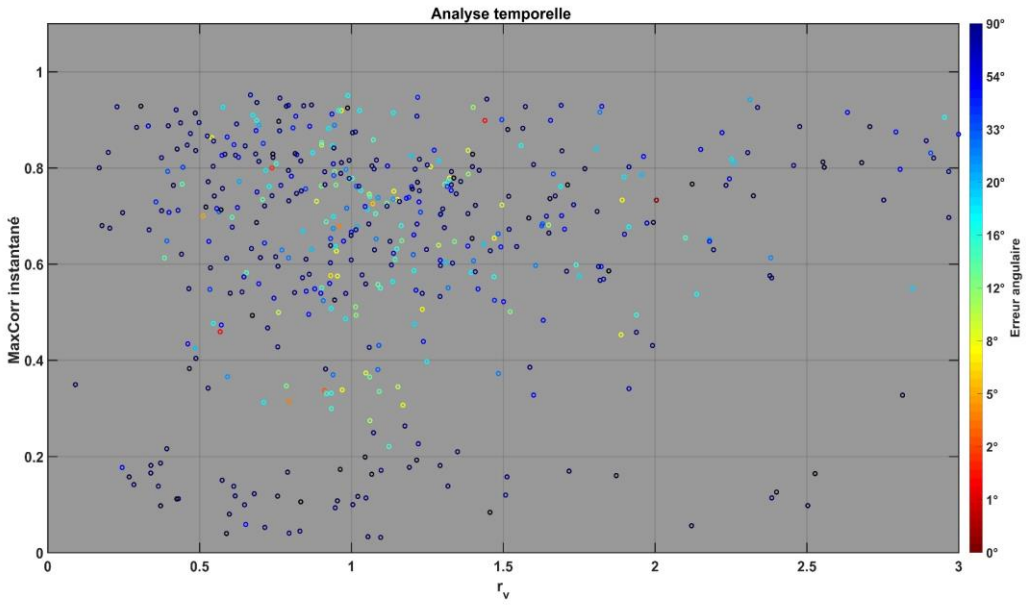

Figure 68. Semblable à la Figure 67 pour la SRIR de la salle Pleyel.

L'impact des ondes secondaires sur la fonction d'intercorrélation et, par conséquent, sur l'estimation de la position montre que de bonnes valeurs se sont concentrées autour de l'amplitude d'intercorrélation 0.8 dans le cas de la SRIR de la salle Varèse (Figure 67). On peut donc supposer que le descripteur « MaxCorr instantané » est moins fiable dans le cas de la réverbération. Dans le cas de la salle Pleyel, les estimations sont plus dispersées par rapport au $\boldsymbol{r_v} = 1$ avec de grandes erreurs (Figure 68). Les perturbations des ondes secondaires dissipent l'énergie de la source acoustique et donc provoquent des erreurs dans l'estimation (Figure 69 et Figure 70). Dans les deux cas on peut remarquer que les points d'estimation ne sont plus concentrés à l'endroit correspondant au $\boldsymbol{r_v} = 1$.

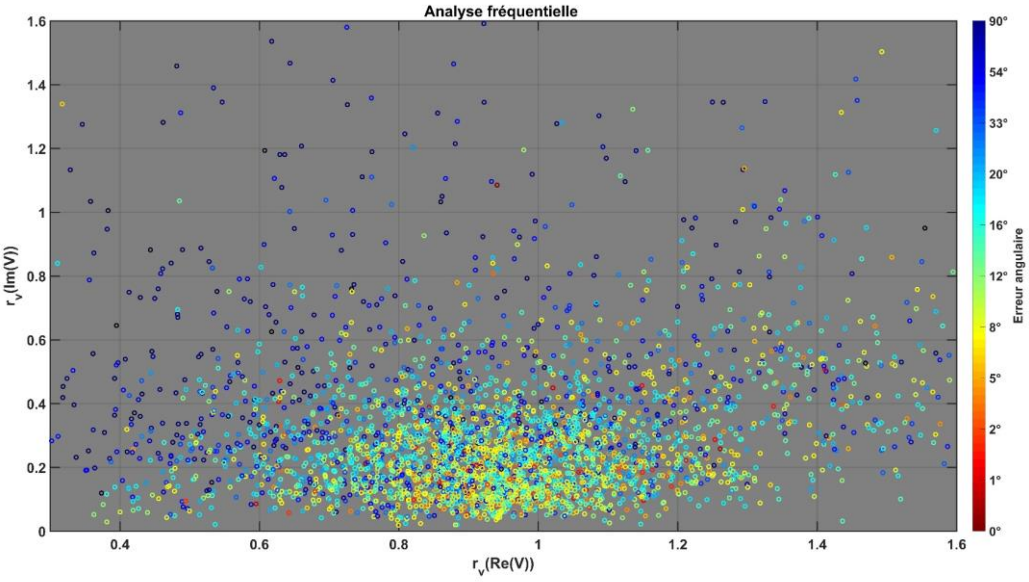

Figure 69. Estimation dans le domaine fréquentiel de la position pour la première source acoustique ('soprano') avec réverbération (SRIR de la salle Varèse) représentée par une erreur angulaire et deux descripteurs : norme de la partie réelle et imaginaire de vecteur vélocité.

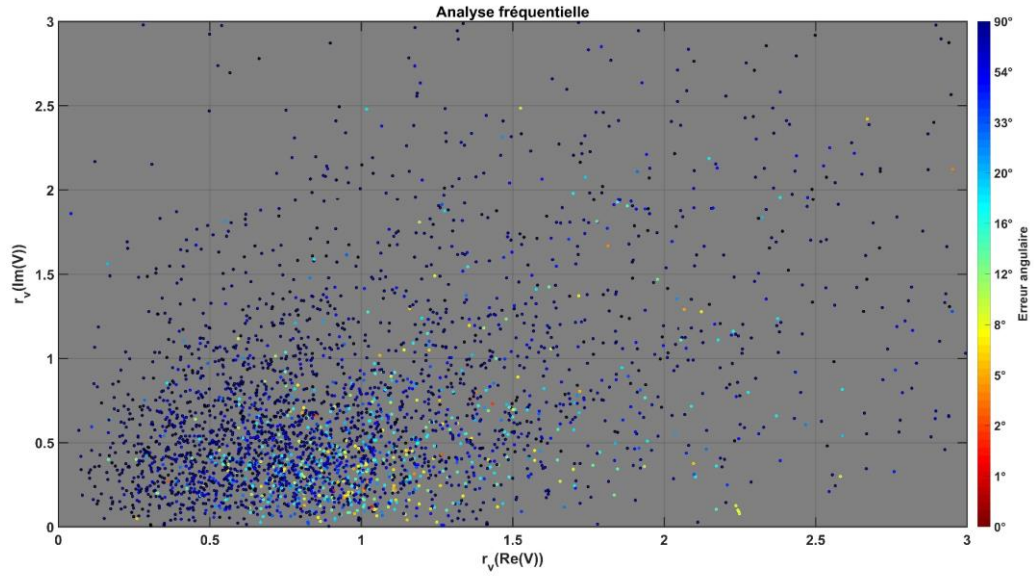

Figure 70. Semblable à la Figure 69 pour la SRIR de la salle Pleyel.

Les représentations temps-fréquence (Figure 71 et Figure 72) montrent bien deux impacts différents de la réverbération sur les estimations pour les deux salles, avec une erreur moyenne ≈ 15° pour la salle Varèse et ≈ 36° pour celle de Pleyel.

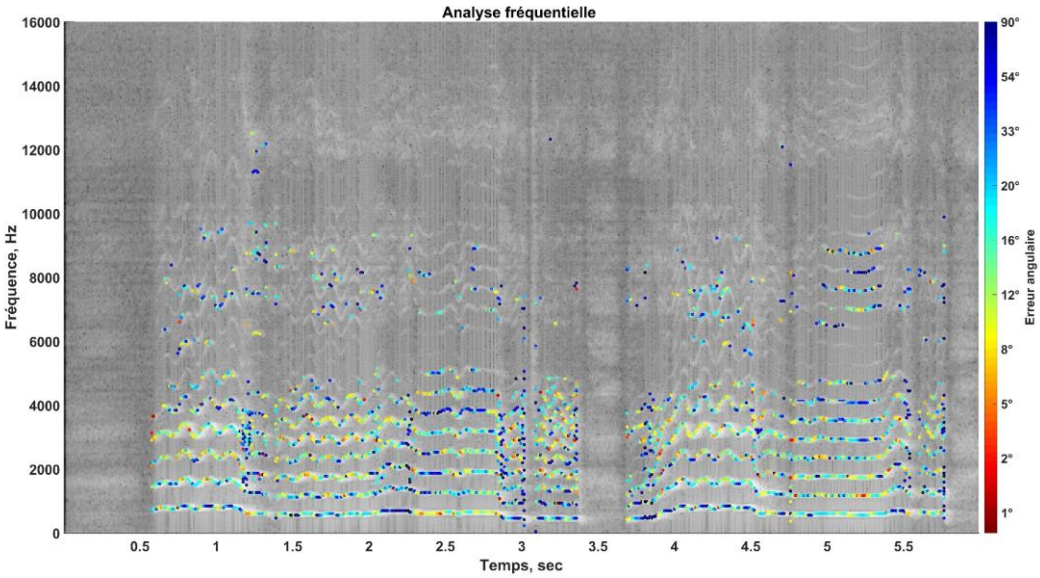

Figure 71. Représentation temps-fréquence de l'erreur angulaire, superposée au spectrogramme de l'extrait analysé de la première source acoustique ('soprano') avec réverbération pour la SRIR de la salle Varèse.

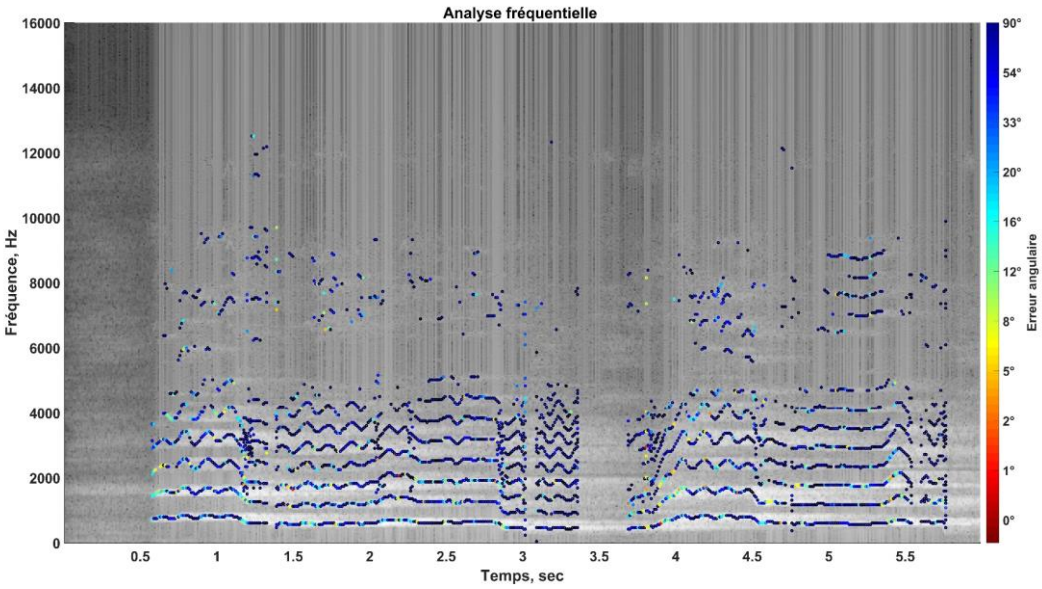

Figure 72. Semblable à la Figure 71 pour la SRIR de la salle Pleyel.

5.2.3 Plusieurs sources acoustiques sans/avec réverbération

Pour cette série de tests la scène sonore reprend la même composition musicale de Mozart, mais avec une formation instrumentale plus complète comprenant les sources acoustiques suivantes :

Sources	Retard (éch.)	Azimut (°)	Elévation (°)	Gain
Soprano	1127	-30	0	1
Basson	1411	45	30	1
Cor d'harmonie	1127	0	0	1
Clarinette	1127	60	0	1
Flûte	4914	90	0	1
Violoncelle	4917	30	0	1
Violon	5077	-60	0	1
Violon alto	4898	-90	0	1
Contrebasse	1127	-45	30	1

Tableau 2. Configuration de la scène sonore simulée. 9 sources acoustiques.

La position et la distance de chaque source acoustique sont choisies arbitrairement en s'inspirant librement des positions typiques d'instruments dans un orchestre symphonique classique (Figure 73). La répartition spatiale est néanmoins déterminée avec la contrainte de

disposer de SRIR correspondant aux angles choisis, pour la version échoïque (ce sont alors les SRIR de la salle Varèse qui sont utilisées).

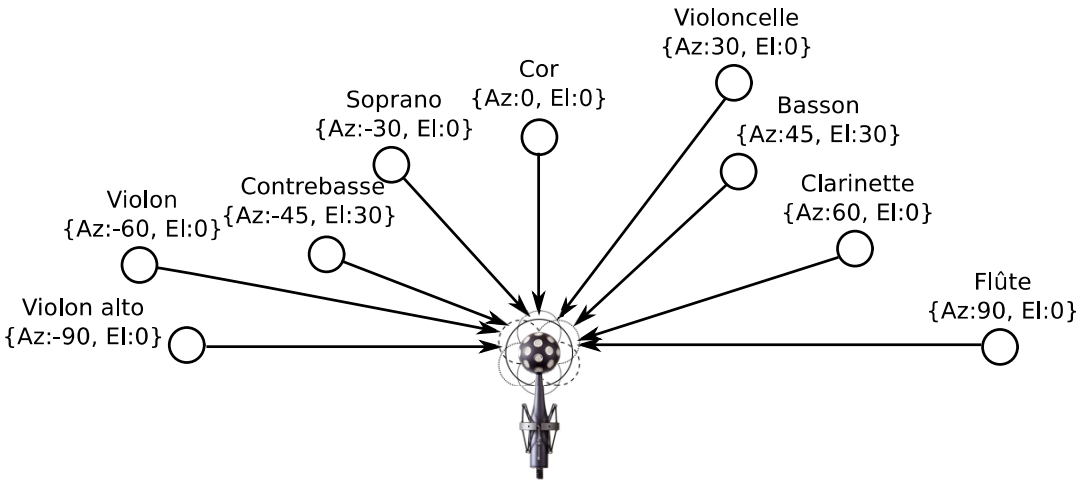

Figure 73. Composition des sources acoustiques dans une scène sonore simulée.

L'idée de ce test est de montrer l'efficacité de l'algorithme dans une scène sonore quasi complète et proche de celle qui serait enregistrée réellement dans un studio d'enregistrement. Pour afficher les résultats d'estimation nous avons choisi la première source acoustique 'soprano' qui fait de la partie des tests dans les sections précédentes et une autre source perturbée, ici 'clarinette'.

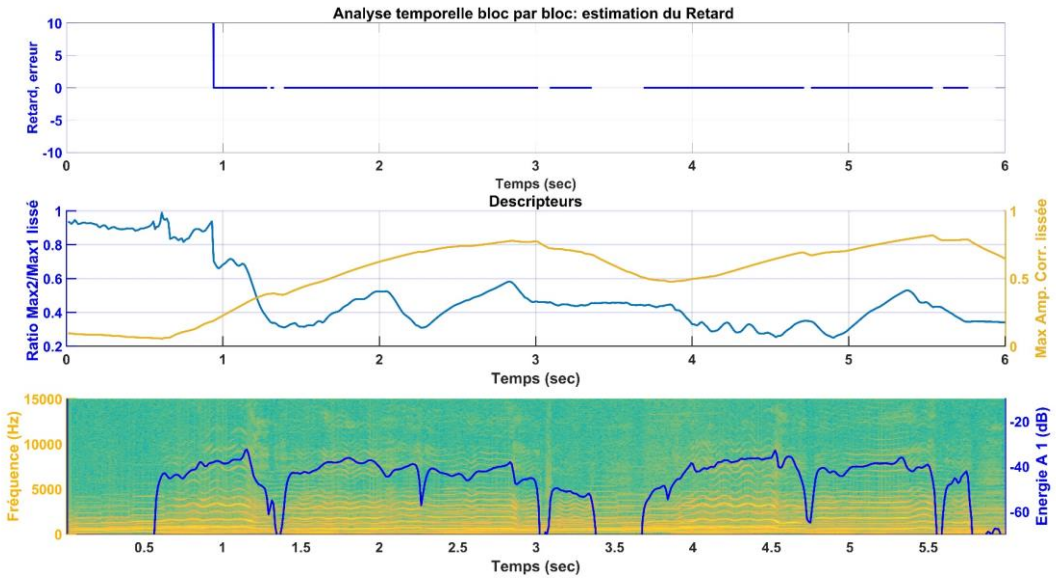

Figure 74. Estimation du retard pour la source acoustique ('soprano') dans la scène sonore composée de 9 sources acoustiques.

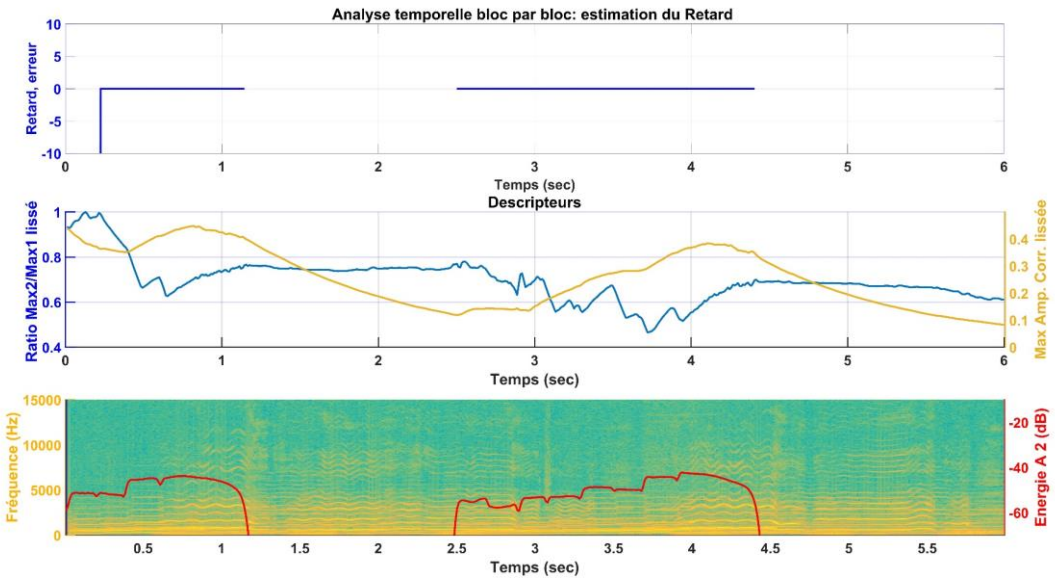

Figure 75. Semblable à la Figure 75 pour la source acoustique 'clarinette'.

L'estimation du retard pour les deux sources acoustiques (Figure 74 et Figure 75) est toujours bonne sauf au début à l'endroit de l'impact de plusieurs sources acoustiques. Le haut niveau de l'amplitude de la fonction d'intercorrélation lissée sur la Figure 74 permet de dire que la source acoustique 'soprano' reste prédominante dans toute la scène sonore. Le ratio des pics maximaux de la fonction d'intercorrélation lissée reste la plupart de temps dans un intervalle de 0,6 à 0,8 à cause de la présence des autres sources. Celles-ci provoquent des erreurs dans l'estimation de la position pour la source 'soprano' au cours de la 3ème seconde (Figure 76) tandis que l'estimation de la position de la source 'clarinette' est plus erronée, surtout en termes d'azimut (Figure 77). Les premières figures suivantes présentent une analyse temporelle de nouveau.

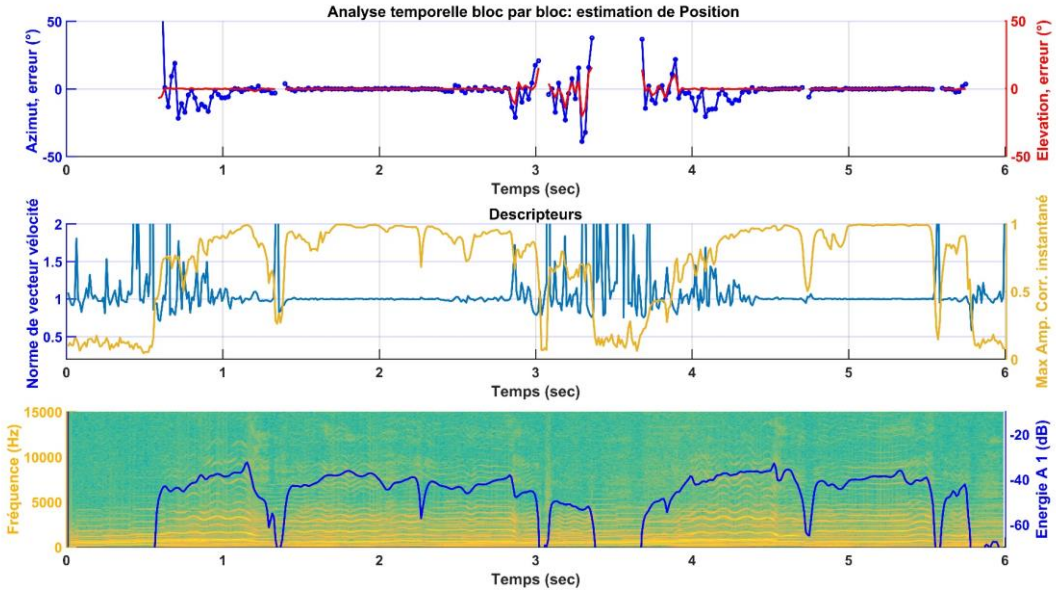

Figure 76. Estimation de l'azimut et d'élévation de la source acoustique ('soprano') dans la scène sonore composée de 9 sources acoustique.

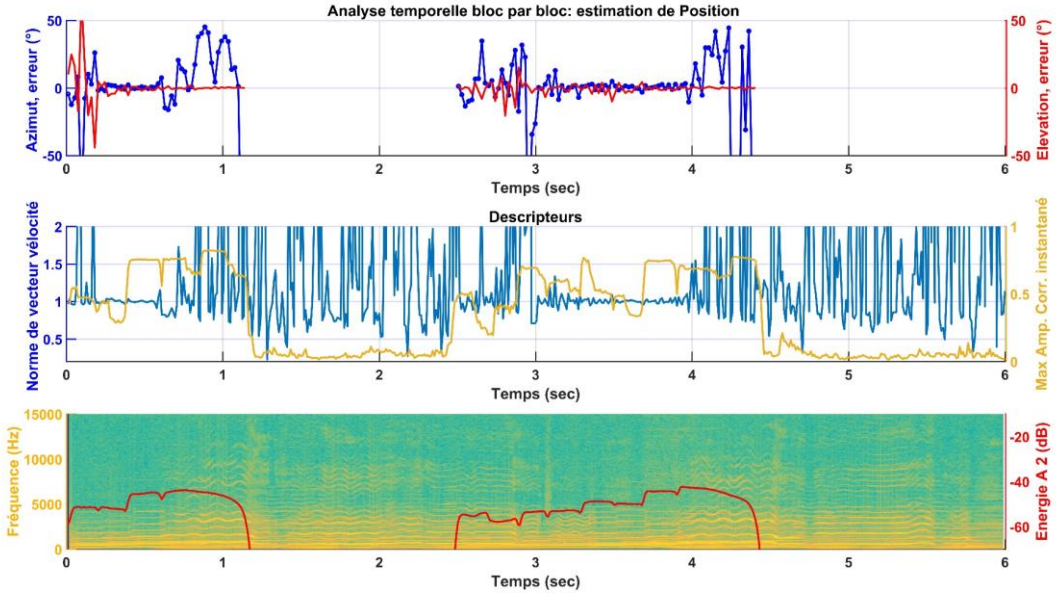

Figure 77. Semblable à la Figure 76 pour la source acoustique 'clarinette'.

Les points d'estimations pour la source 'soprano' gardent le même caractère de dispersion, quant au descripteur $\boldsymbol{r_v}$ (Figure 78), que dans les sections précédentes du chapitre ce qui se traduit par la plage de valeurs de $\boldsymbol{r_v}$ plus large, comprise entre 0,5 et 1,5 , une estimation correcte étant toujours associée à un indice $\boldsymbol{r_v}$ est proche de 1.

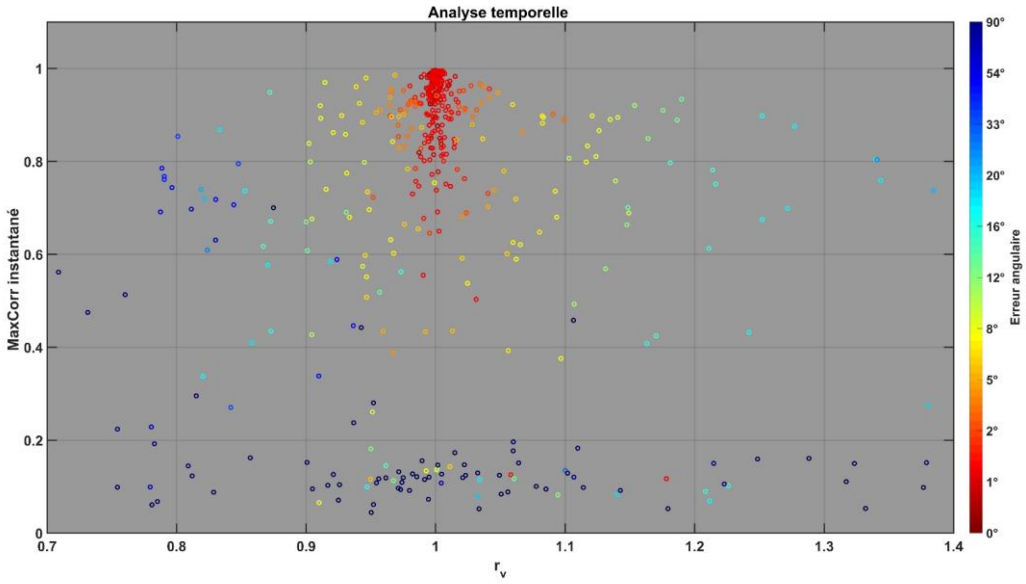

Figure 78. Estimation dans le domaine temporel de la position pour la source acoustique ('soprano') avec une erreur angulaire et deux descripteurs : norme du vecteur vélocité « r_v » et amplitude maximale de la fonction d'intercorrélation lissée « MaxCorr ».

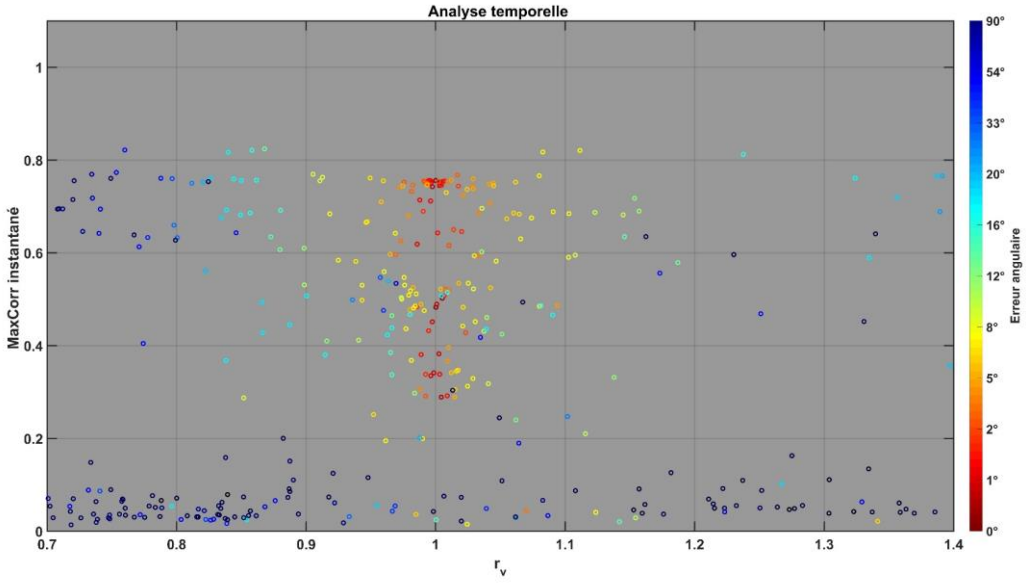

Figure 79. Semblable à la Figure 78 pour la source acoustique 'clarinette'.

Si l'on applique maintenant une analyse fréquentielle, on observe que la représentation temps-fréquence de l'erreur angulaire (Figure 80 et Figure 81) est peu changée, dans cette composition spatiale plus complexe, comparée au cas avec deux sources acoustiques (voir section 5.2.1). Il y a néanmoins certains endroits (comme au milieu) où l'on observe une

dégradation légère de l'estimation, pour des fréquences partagées par plusieurs sources simultanément.

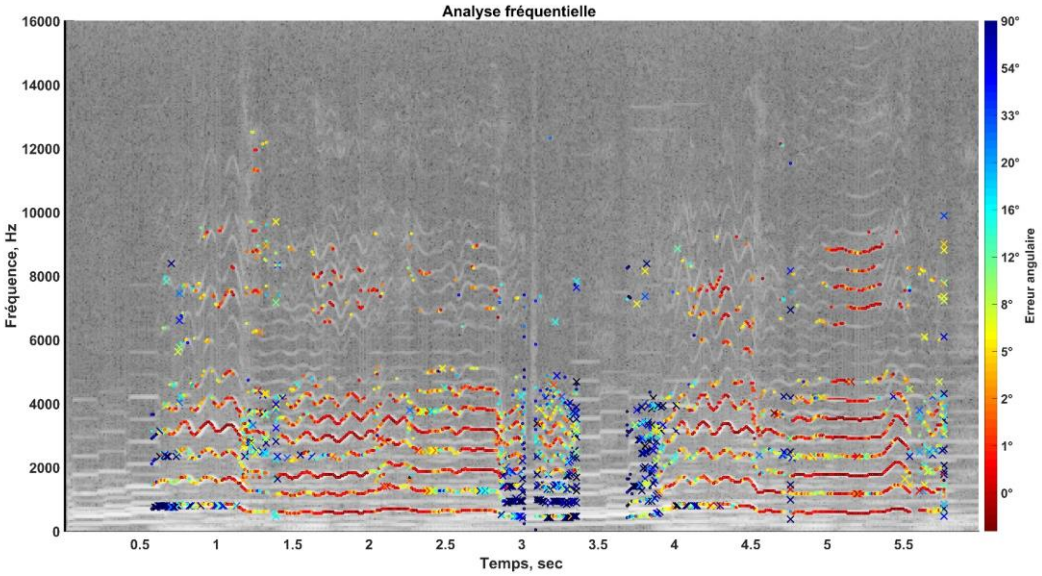

Figure 80. Représentation temps-fréquence de l'erreur angulaire, superposée au spectrogramme de l'extrait analysé de la source acoustique ('soprano') pour chaque bin fréquentiel interféré (« croix ») et non-interféré (« point »).

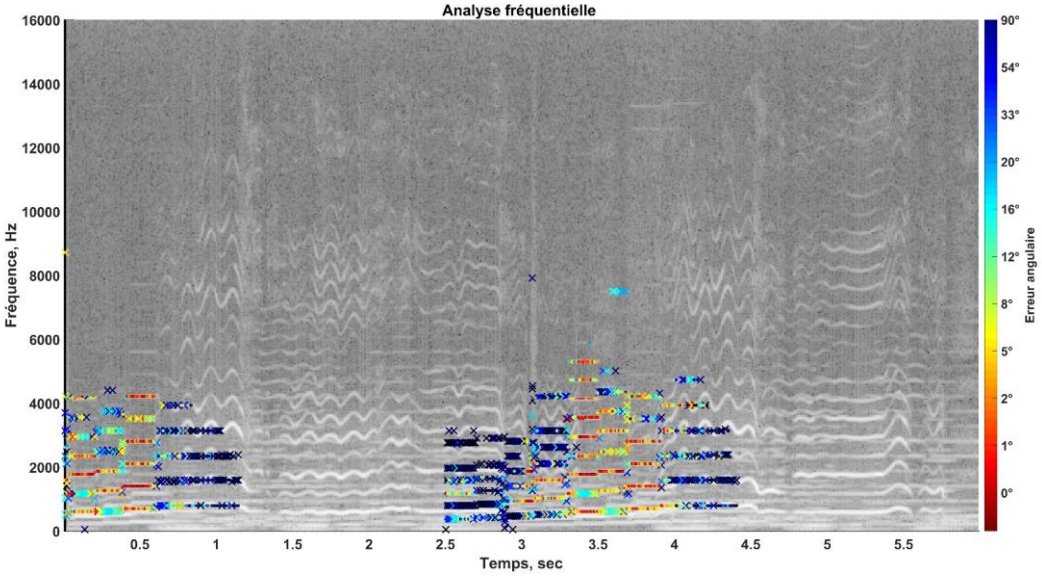

Figure 81. Semblable à la Figure 84 pour la source acoustique ('clarinette').

Considérons maintenant la même composition sonore en présence d'effet de salle, celui-ci étant simulé en utilisant les SRIR de la salle Varese associées aux mêmes positions de sources acoustiques. On constate que l'estimation du retard n'est pas changée et reste fiable la plupart de temps (Figure 82 et Figure 83). La représentation temps-fréquence (Figure 84 et Figure 85)

montre que l'estimation pour la source 'soprano' est moins erronée, par rapport à celle de la source 'clarinette'. Pour la source 'soprano' on obtient des erreurs faibles surtout sur les harmoniques élevés et du vibrato qui sont probablement liées au fait que les fréquences changent vite et le son direct se mélange moins avec le son indirect. Une possibilité d'obtenir une estimation plus précise est de prendre en compte les observations sur l'ensemble des trames et d'en dégager une moyenne statistique pondérée par un indice de confiance.

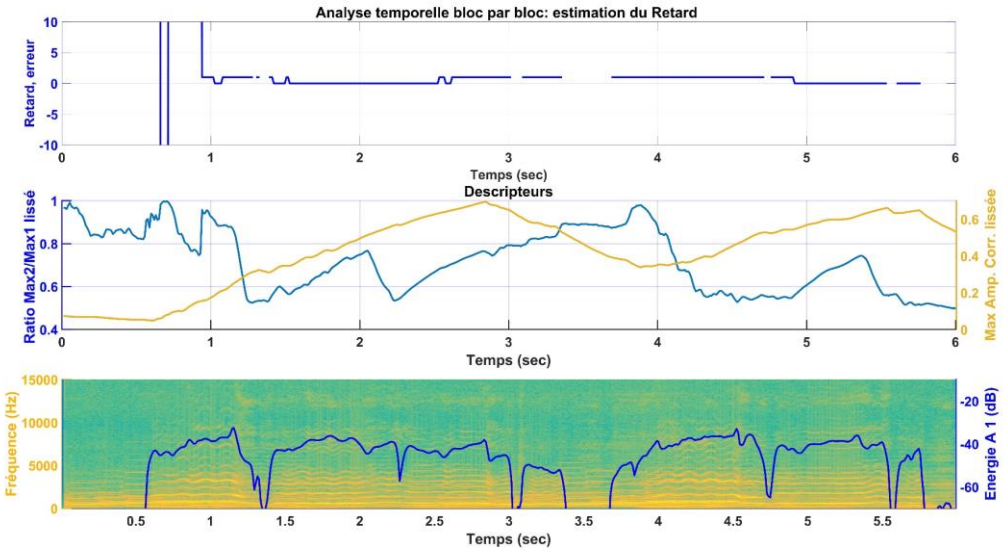

Figure 82. Estimation du retard pour la première source acoustique ('soprano') dans la scène sonore composée de 9 sources acoustiques avec réverbération.

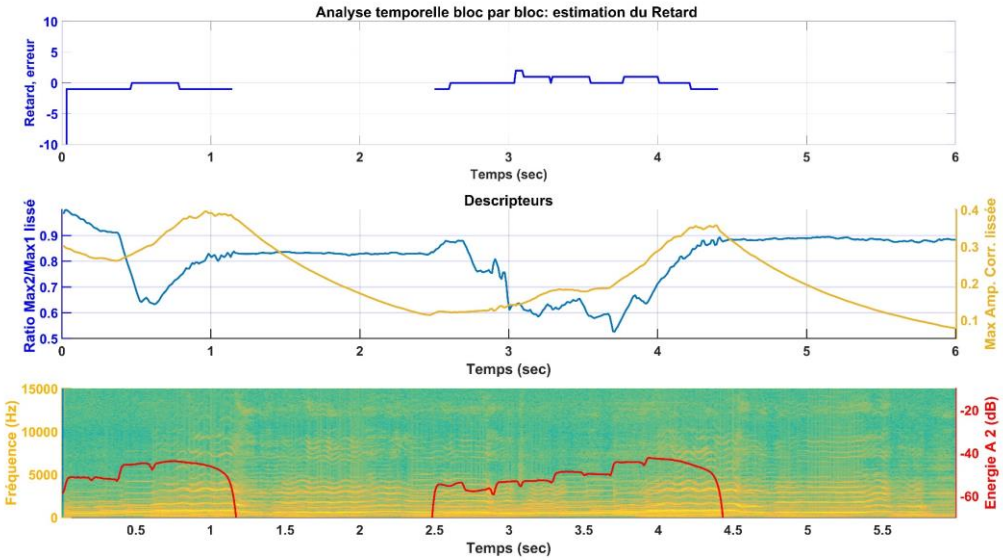

Figure 83. Semblable à la Figure 82 pour la source acoustique 'clarinette'

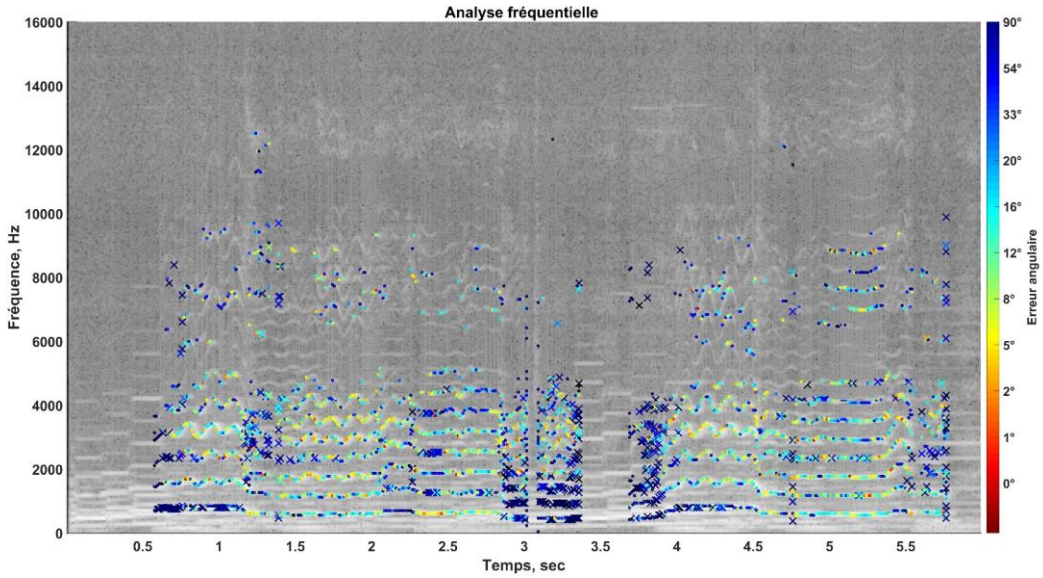

Figure 84. Représentation temps-fréquence de l'erreur angulaire, superposée au spectrogramme de l'extrait analysé de la source acoustique ('soprano') pour chaque bin fréquentiel non-interféré.

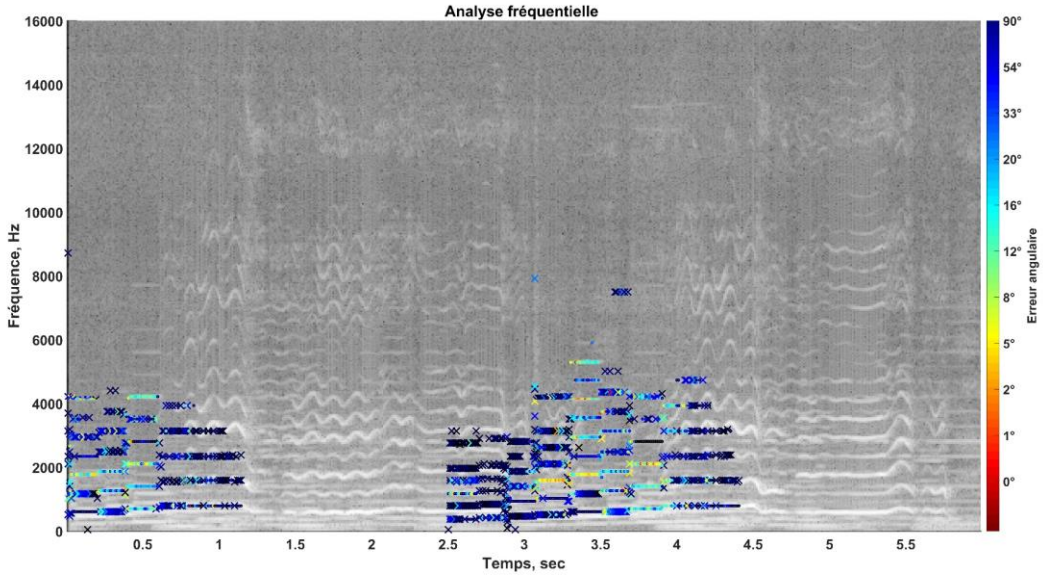

Figure 85. Idem pour la source acoustique 'clarinette'

5.3 Scène sonore réelle

Dans le cadre de cette thèse une prise de son HOA a été effectuée au sein de Conservatoire National Supérieur de Musique et de Danse de Paris (CNSMDP). Le microphone principal (Eigenmike) était suspendu au-dessus du pupitre du chef d'orchestre (Figure 86).

Figure 86. Disposition du microphone principal pendant la prise de son au CNSMDP

Les positions des musiciens décrivent un demi-cercle par rapport au pupitre du chef d'orchestre (Figure 87). Chaque microphone d'appoint capte également d'autres sources acoustiques dans une salle réverbérante, on est donc en situation de diaphonie. Le microphone principal capte des ondes acoustiques secondaires dont les plus précoces proviennent du sol et du pupitre.

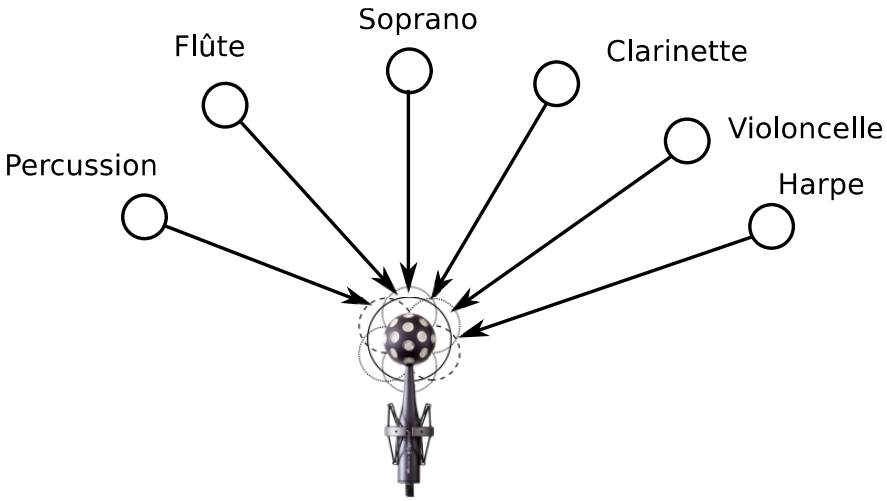

Figure 87. Position approximative de chaque source acoustique pendant la prise de son HOA au CNSMDP.

On observe que l'estimation du retard pour la première source acoustique (« soprano ») est moins fiable au cours du temps à cause de la contribution deuxième source acoustique « flûte » qui joue en même temps (Figure 88).

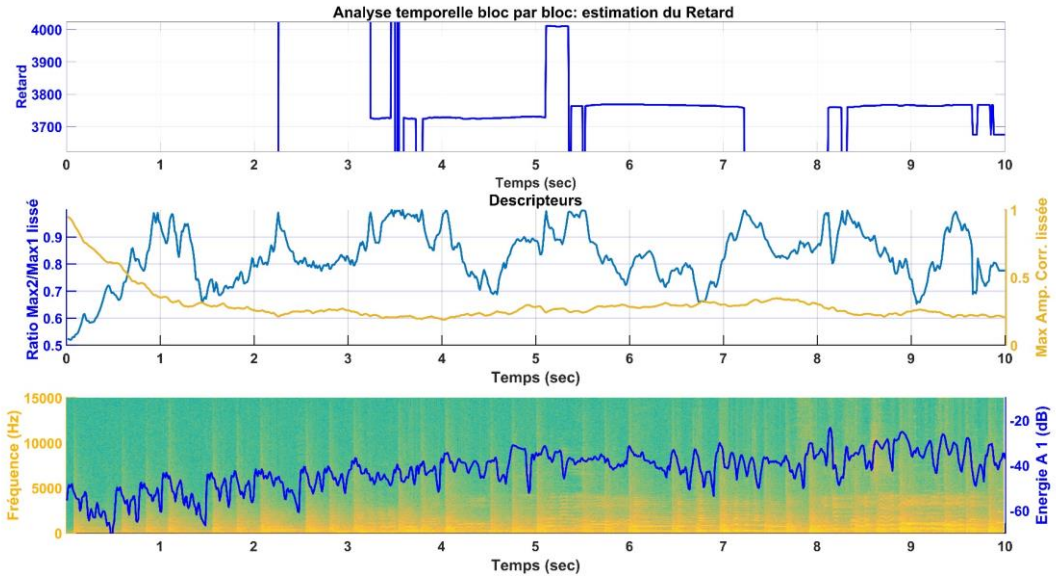

Figure 88. Estimation du retard pour le signal du microphone d'appoint à côté de « soprano » d'une scène sonore enregistrée au CNSMDP.

L'amplitude de la fonction d'intercorrélation reste faible la plupart de temps à cause de plusieurs facteurs tels que la réverbération de la salle, la diaphonie entre les sources acoustiques concurrentes et le microphone d'appoint dédié à la voix. Aux endroits avec des perturbations d'estimation du retard (par exemple $3^{ème}$-$4^{ème}$ et $5^{ème}$-$6^{ème}$ secondes) on peut constater que des pics maximaux de la fonction d'intercorrélation sont proches. Donc le ratio « Ratio Max2/Max1 » dans ces cas est proche de 1 et indique les endroits où la fonction d'intercorrélation peut se tromper avec la détection du pic maximal et, par conséquent, avec l'estimation du retard. En revanche, avec le ratio faible on observe une plage de valeurs fixes dans un intervalle 3700-3800 échantillons et on peut donc supposer que le retard réel pour cette source acoustique se trouve dans cet intervalle.

Dans le domaine fréquentiel l'estimation de l'azimut et de l'élévation est faite pour chaque bloc de référence pour les fréquences non-interférées. La position de la première source acoustique (« soprano ») paraît mieux prédite par l'analyse fréquentielle que par l'analyse temporelle (Figure 89). En effet la dispersion de l'histogramme est clairement moindre.

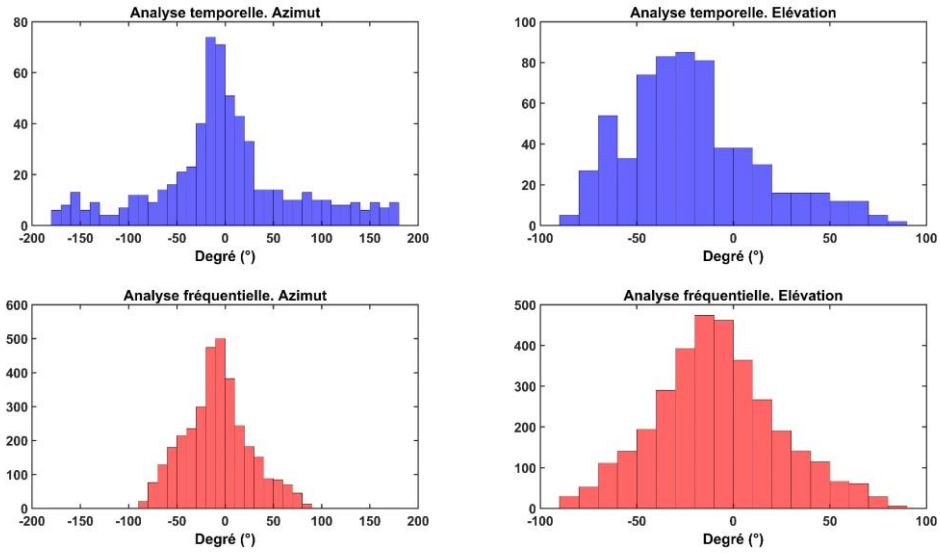

Figure 89. Estimation de l'azimut et de l'élévation dans le domaine temporel (bleu) et fréquentiel (rouge) à partir du signal de microphone d'appoint à côté d'une source acoustique « soprano ».

Afin d'améliorer les résultats lorsqu'on applique l'analyse temporelle, il est peut être judicieux d'augmenter la taille de bloc de référence (sur lequel est calculé la corrélation). En effet, des contributions distinctes présentent en général un indice de corrélation d'autant plus faible que celui-ci est calculé sur le long terme, donc une taille de bloc plus grande tend à gommer les perturbations occasionnées par les autres sources. Par exemple, pour la source « soprano » on obtient des résultat plus fiables et moins dispersés, dans le domaine temporel (Figure 90).

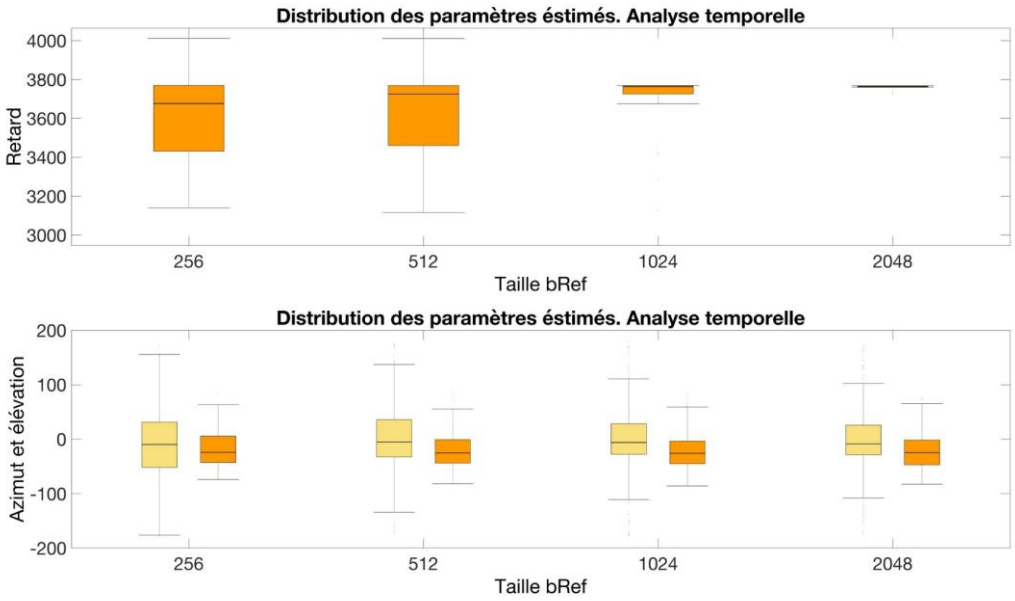

Figure 90. Boîte à moustache d'estimation du retard, de l'azimut et d'élévation dans le domaine temporel de la première source acoustique « soprano » pour les blocs de références de tailles différentes.

La deuxième source (« flûte ») que nous avons choisie pour l'analyse est à proximité d'une percussion et joue simultanément avec celle-ci (Figure 91).

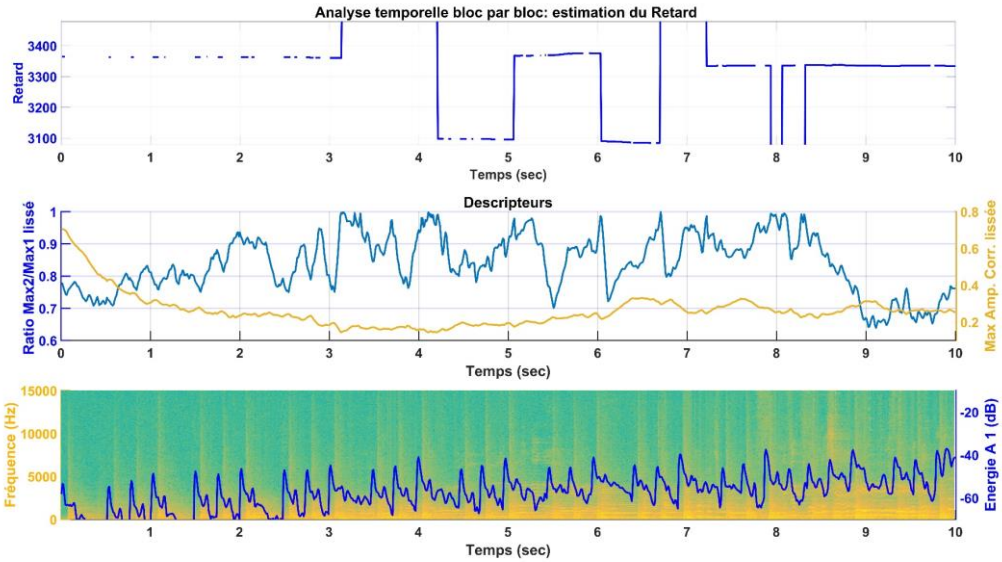

Figure 91. Estimation du retard pour le signal du microphone d'appoint à côté de « flûte » d'une scène sonore enregistrée au CNSMDP.

A partir de l'estimation présentée sur la Figure 91 on peut supposer que le retard du signal du microphone d'appoint associé à la source « flûte » par rapport au signal du microphone principal est situé dans un intervalle 3300-3400 échantillons. On peut remarquer

aussi des fortes perturbations de l'estimation au moment de l'impact d'une première partie de la source « soprano » entre la 3ème et la 5ème seconde et celui des autres sources acoustiques entre la 6ème et la 7ème seconde. Avec l'augmentation du niveau sonore de la source « flûte » vers la fin du morceau l'estimation se stabilise vers une valeur fixe avec le ratio faible des pics maximaux de la fonction d'intercorrélation.

La position angulaire approximative de la source « flûte » (Figure 86 et Figure 87) est environ de 0° en l'élévation et de -70° en l'azimut. L'estimation de ces paramètres dans le domaine temporel ne peut être pas considérée comme fiable à partir de l'histogramme pour l'élévation et l'azimut qui est largement dispersé (Figure 92, histogramme en bleu). En revanche, la position paraît mieux prédite dans le domaine fréquentiel (Figure 92, histogramme en rouge) où on peut distinguer une valeur de -70° la plus fréquemment estimée pour l'azimut et de -10° pour l'élévation.

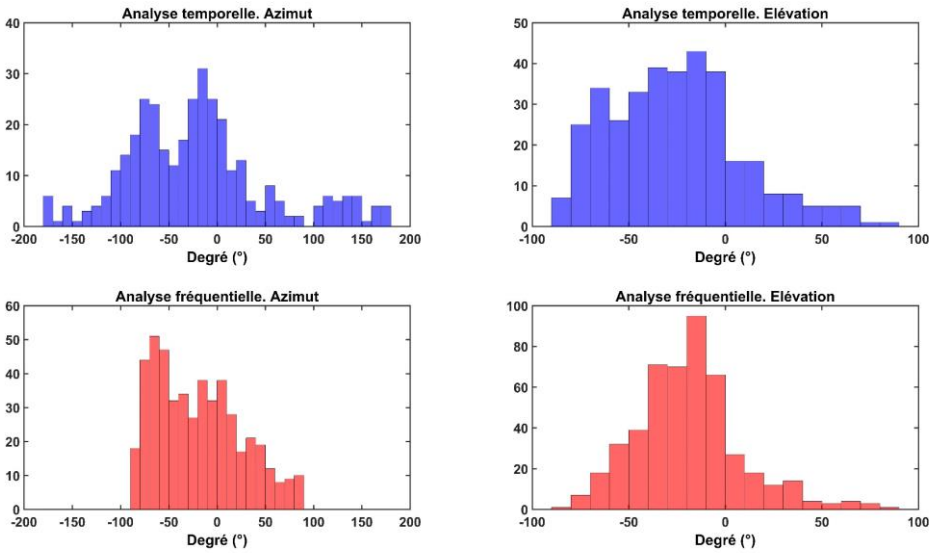

Figure 92. Estimation de l'azimut et de l'élévation dans le domaine temporel (bleu) et fréquentiel (rouge) à partir du signal de microphone d'appoint à côté d'une source acoustique « flûte ».

Dans ce cas réel en absence de l'information exacte sur la position de chaque source acoustique on ne peut pas estimer l'erreur angulaire. En revanche, en utilisant le descripteur $\boldsymbol{r_v}$ et en prenant en compte que pour de bonnes estimations $\boldsymbol{r_v}$ proche de 1 on peut observer la dispersion des estimations de la position dans le plan azimut-élévation. Pour la première source acoustique (Figure 93) on observe un nuage des estimations avec le $\boldsymbol{r_v}$ entre 0,8 et 1,2 concentré autour de -10° pour l'azimut et l'élévation qui n'est pas loin d'une valeur attendue de 0°. Mais on constate la difficulté de prédiction de la position causée par une dispersion forte des autres

estimations avec le $\boldsymbol{r_v}$ proche de 1 par rapport aux valeurs attendues. Malgré la dispersion forte des estimations pour la deuxième source « flûte » on distingue mieux le nuage des points avec $\boldsymbol{r_v}$ proche de 1 (Figure 94) concentré autour de -70° pour l'azimut et -25° pour l'élévation.

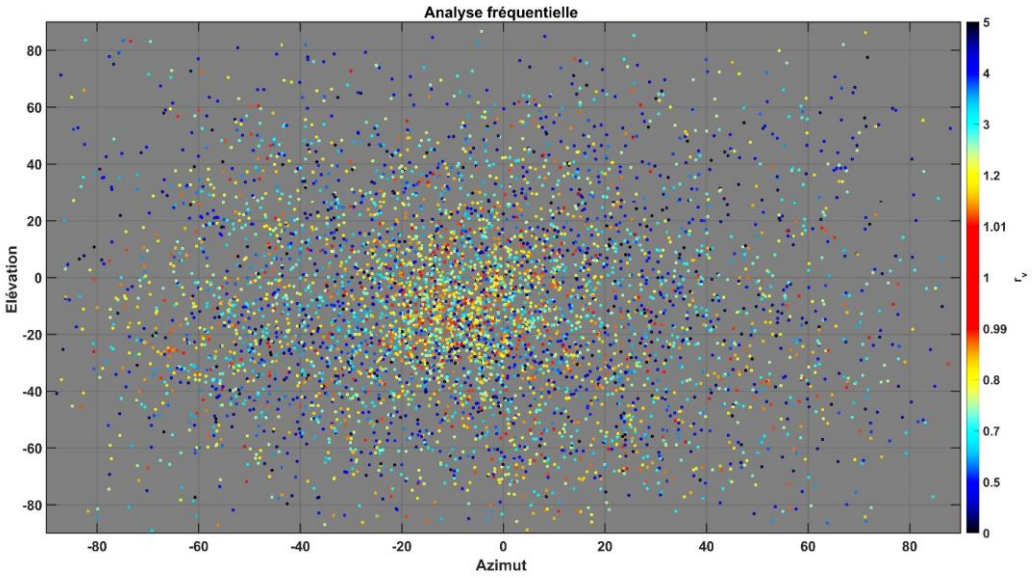

Figure 93. Estimation de la position dans le domaine fréquentiel pour les fréquences interférées par rapport au descripteur r_v pour la première source acoustique « soprano »

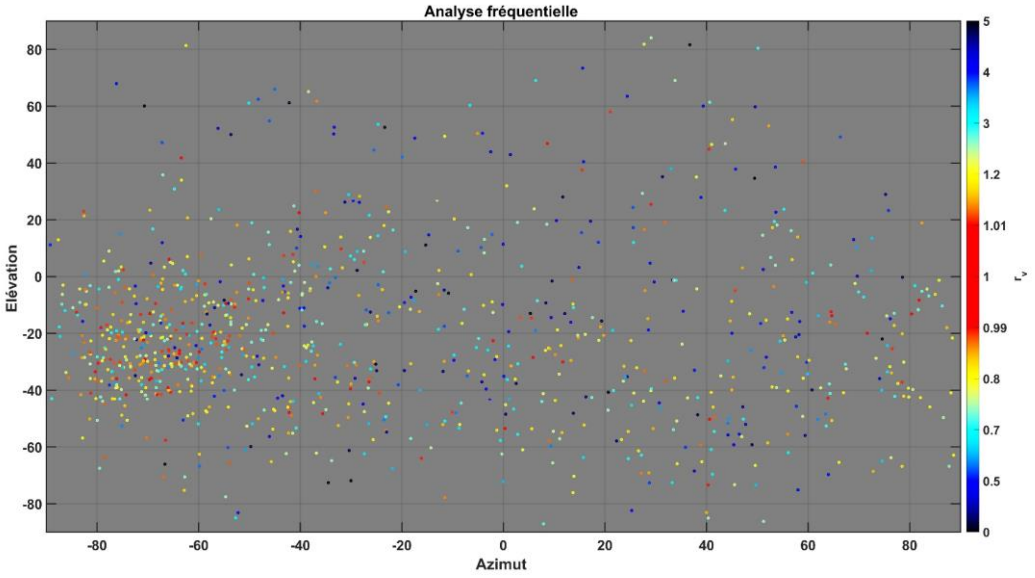

Figure 94. Semblable à la Figure 93 pour la deuxième source acoustique « flûte ».

Dans le cas réel et complexe on peut conclure que l'utilisation de la fonction d'intercorrélation lissée paraît fiable pour la prédiction du retard du signal de chaque microphone d'appoint par rapport au signal du microphone principal. En revanche, avec les facteurs tels que la réverbération de la salle, la diaphonie entre les sources acoustiques

concurrentes, l'estimation de la position basée sur le descripteur $\boldsymbol{r_v}$ est plus pertinente dans le domaine fréquentiel.

5.4 Conclusion

Dans ce chapitre nous avons effectué des tests avec des scènes sonores simulées et une scène sonore réelle complexe enregistrée au CNSMDP à Paris. Nous avons montré une complexité croissante des scènes sonores simulées à partir du cas simple composé de deux sources acoustiques et du cas plus complexe avec plusieurs sources acoustiques et réverbération.

L'estimation des paramètres dans le domaine temporel dans une scène sonore simple avec deux sources acoustiques est robuste sauf en quelques endroits où sont présentes de petites perturbations. Nous avons testé plusieurs descripteurs afin de montrer l'efficacité de l'algorithme ainsi que des difficultés rencontrées pendant l'estimation. Pour augmenter la robustesse de l'algorithme dans le domaine fréquentiel, la méthode d'extraction des fréquences non-interférées a été appliquée pendant l'analyse. Ainsi nous avons proposé une méthode d'affichage des résultats basée sur l'erreur angulaire**.** La norme du vecteur vélocité en tant que descripteur le plus promettant a été étudiée dans tous les cas simulés. La réverbération introduite dans le système montre que l'estimation de la position devient plus difficile et elle est souvent erronée tandis que l'estimation du retard est toujours bonne grâce à la fonction d'intercorrélation lissée.

Ainsi nous avons étudié une scène sonore complexe avec 9 sources acoustiques sans et avec réverbération. En ce qui concerne de l'estimation du retard, l'algorithme est toujours robuste en utilisant la fonction d'intercorrélation lissée. Dans le cas sans réverbération l'estimation de la position des sources acoustiques analysées dans le domaine temporel est un peu erronée et donc nous avons étudié l'estimation dans le domaine fréquentiel pour diminuer la variation d'erreurs.

Le cas réel à partir de la prise de son HOA effectuée au CNSMDP montre toutes les difficultés évoquées pendant les tests avec les scènes sonores simulées. La méthode d'analyse du retard à l'aide de la fonction d'intercorrélation lissée montre de bonnes estimations. Mais l'estimation de l'azimut et de l'élévation est plus dispersée par rapport aux valeurs attendues. Une amélioration a été remarquée après l'utilisation des méthodes d'analyse dans le domaine fréquentiel. Par contre, il est utile d'étudier des méthodes d'estimation avec facteur d'oubli adaptatif et développer l'analyse dans le domaine fréquentiel. Ainsi il manque des tests

subjectifs pour trouver le seuil des erreurs angulaires audibles à l'oreille humaine pendant le mixage.

6 Conclusion globale

Le travail de cette thèse est associé à l'assistance au mixage entre un microphone principal de type HOA (« Higher Order Ambisonics ») et des microphones d'appoint, afin de faciliter le travail de l'ingénieur du son pendant le mixage.

Nous avons étudié le concept de la prise du son classique et du mixage stéréo ainsi que les problèmes souvent rencontrés pendant le mixage, telles que les mesures « à la main » (avec un mètre par exemple) ou « à l'oreille » (en agissant sur le retard du son) de la distance entre les microphones et l'identification du positionnement spatial des sources acoustiques. Nous avons montré le concept de la prise de son immersive à base de signaux Ambisonic d'ordre 1 et abordé le mixage du son immersif avec le cas particulier du format HOA. Dans la cadre de cette thèse nous avons effectué plusieurs enregistrements avec un microphone Ambisonic afin de maîtriser la prise de son HOA et de générer du contenu pour effectuer des tests sur les algorithmes développés.

Après avoir montré la problématique globale du mixage HOA avec microphones d'appoint, nous avons présenté sous certaines hypothèses une formalisation mathématique du problème. A partir de ce formalisme, deux étapes d'estimation des paramètres de mixage ont été abordées : identification du retard dans le domaine temporel puis estimation de l'azimut et de l'élévation (dans le domaine temporel mais aussi fréquentiel) à partir du retard estimé. Sur la base d'un cas simple avec une source acoustique et un microphone d'appoint associé nous avons évoqué l'identification du retard basé sur la fonction d'intercorrélation entre le signal du microphone d'appoint et celui du microphone principal (première composante Ambisonic d'ordre 1). Nous avons montré que la fonction d'intercorrélation classique provoquait des problèmes de robustesse dans les cas avec plusieurs sources acoustiques dans les mêmes bandes de fréquences, avec réverbération ou dans le cas d'un signal périodique. Pour cette raison nous avons proposé une extension à la fonction d'intercorrélation classique qui est capable de mémoriser et cumuler les fonctions d'intercorrélations calculées avec un certain poids. Cette extension (nommée « intercorrélation lissée », qui a fait l'objet d'une demande de brevet [2]) montre des résultats plus fiables dans les cas évoqués. Une autre méthode de l'identification du retard basée sur les spectrogrammes des signaux et l'intercorrélation généralisée n'est pas abordée dans cette thèse et peut être trouvée dans le travail [28].

Sous forme analytique dans le domaine temporel, nous avons déduit l'expression de l'azimut et l'élévation de la source à partir des signaux du microphone principal (Ambisonic d'ordre 1) et du microphone d'appoint. La robustesse de ces méthodes d'estimation a fait l'objet

d'études dans le domaine fréquentiel. Nous avons évoqué l'estimation de l'azimut et de l'élévation pour chaque bin fréquentiel et développé une analyse fréquentielle dans le cas de plusieurs sources acoustiques afin de déterminer comment les sources interfèrent.

Une autre contribution de cette thèse a été consacrée aux « descripteurs » qui ont été introduits pour améliorer la fiabilité des estimations (ce qui a aussi fait l'objet d'une demande de brevet [2]). Une partie des descripteurs proposés évoque une caractéristique sonore (l'énergie du signal de microphone d'appoint, l'amplitude et le rapport entre les pic maximaux d'intercorrélation, etc.) afin de se focaliser dans un endroit significatif du signal pendant l'analyse. L'autre partie des descripteurs est liée à la caractéristique spatiale de la source (le vecteur vélocité et sa norme). La norme du vecteur vélocité étant un descripteur potentiellement fiable, elle a fait l'objet d'une attention particulière dans cette thèse. Différentes combinaisons de descripteurs peuvent former les « indices de confiance » qui ouvrent une piste de recherches et qui peuvent potentiellement améliorer la robustesse des estimations effectuées. Dans ce domaine on peut citer un travail récent qui introduit l'indice de confiance pour justifier le choix du retard entre les deux signaux basé sur la fonction d'intercorrélation [29].

Afin d'anticiper l'intégration des techniques développées dans un logiciel (« plugin ») audio nous avons conçu l'algorithme d'analyse des paramètres. Nous avons montré l'utilisation des descripteurs et des indices de confiance à l'intérieur de l'algorithme. A partir des paramètres estimés, nous avons expliqué l'étape finale du mixage HOA. L'algorithme, implémenté sous Matlab, peut être considéré comme un prototype du plugin audio d'assistance au mixage HOA. Une partie du programme consiste en la création de la scène sonore à partir des signaux mono ou à partir d'un enregistrement HOA, mais aussi l'analyse dans le domaine temporel ou fréquentiel, le mixage HOA final avec les signaux des microphones d'appoint ajustés avec les paramètres de mixage estimés et à la restitution du mixage final vers différents systèmes de diffusion du son (binaural, 5.1 etc).

Pour valider l'algorithme nous avons effectué plusieurs tests sous Matlab. Pour une série de tests nous avons utilisé des scènes simulées à partir de sources sonores au format mono. Un module de programme permet de positionner spatialement chaque source acoustique et de créer la scène sonore selon plusieurs conditions telles que réverbération, diaphonie et permet également d'effectuer l'encodage spatial au format HOA. L'analyse des paramètres de mixage a montré que le retard était bien estimé en utilisant la fonction d'intercorrélation lissée dans les cas simples (deux sources acoustiques) et complexes (plusieurs sources acoustiques, avec réverbération ou diaphonie). L'estimation de l'azimut et de l'élévation est moins fiable dans le domaine temporel surtout s'il y a un certain niveau de réverbération. L'analyse dans le domaine

fréquentiel montre une amélioration pour l'estimation de la position mais nous avons remarqué des erreurs dans plusieurs cas. Des descripteurs et des indices de confiance sont proposés afin qu'un module, qu'il reste à concevoir, puisse fournir une bonne estimation parmi toutes les estimations proposées au cours du temps. Une autre série de tests a été effectuée à partir d'un enregistrement réel réalisé dans la cadre de cette thèse avec le microphone Ambisonic Eigenmike (scène constituée par 9 sources acoustiques captées également par microphones d'appoint associés dans une salle de concert du CNSDMP à Paris). L'algorithme montre alors une bonne robustesse de l'estimation du retard, par contre la position estimée de chaque source acoustique reste moins précise.

Des tests subjectifs permettant de valider les résultats obtenus lors de ce travail n'ont pas eu le temps d'être menés. Une piste de recherche pour de futurs travaux repose sur le choix et l'utilisation de descripteurs et d'indices de confiances pouvant améliorer les résultats actuels, à la fois dans le domaine temporel et fréquentiel. La norme du vecteur vélocité a fait l'objet d'une recherche particulière en tant que descripteur, et a montré des perspectives intéressantes pour l'estimation de la position.

Pour conclure, nous pouvons remarquer que depuis une dizaine d'années, l'intérêt pour la technologie HOA progresse, avec notamment son support par le nouveau codec « MPEG-H 3D Audio ». Les microphones Ambisonic deviennent financièrement plus accessibles et l'enregistrement HOA va être probablement intégré nativement dans la chaîne de production audio au cours de la prochaine décennie. Dans ce contexte il y a nécessité d'avoir à disposition des outils au format HOA, qui sont peu nombreux sur le marché aujourd'hui. Le mixage au format HOA avec les signaux de microphones d'appoint est une étape dans le développement des outils classiques dans le domaine HOA et ce sujet mérite de faire l'objet d'une recherche complémentaire.

Annexe 1 : Captation et mixage du son

Stéréo

Pour localiser une source acoustique les oreilles humaines utilisent les indices de localisation basés sur des différences de temps ΔT (**I**nteraural **T**ime **D**ifferences - ITD) et de niveau ΔI (**I**nteraural **L**evel **D**ifferences - ILD) entre les 2 oreilles. Dans une restitution sur une paire de haut-parleurs, chaque oreille reçoit un mélange de deux signaux incluant un retard et une atténuation [6]. L'**ITD** ou différence interaurale de temps d'arrivée de l'onde sonore entre les deux oreilles est l'indice de localisation prédominant pour les basses fréquences. Par ailleurs, pour les hautes fréquences, l'indice de localisation prépondérant correspond à la différence interaurale d'amplitude de l'onde sonore entre les deux oreilles (appelé « **ILD ») (Figure 95) [30].

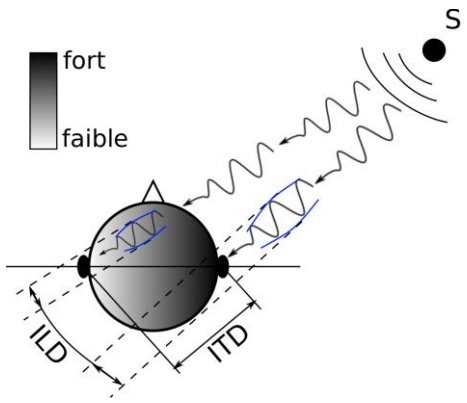

Figure 95. Les indices de latéralisation. ITD et ILD.

Microphone et directivité

Le microphone est un appareil qui est capable de transformer l'énergie mécanique en énergie électrique. Malgré la diversité des microphones, ils sont essentiellement basés sur le principe de la vibration d'une membrane. La vibration des particules dans l'air (Figure 96) engendre des ondes acoustiques, nommées incidentes, qui à leur tour, à proximité d'une membrane qui est fixée dans le corps du microphone, provoquent la vibration mécanique de celle-ci. La vibration de la membrane est responsable de l'apparition d'un courant électrique ou d'une tension grâce à la différence de potentiel générée.

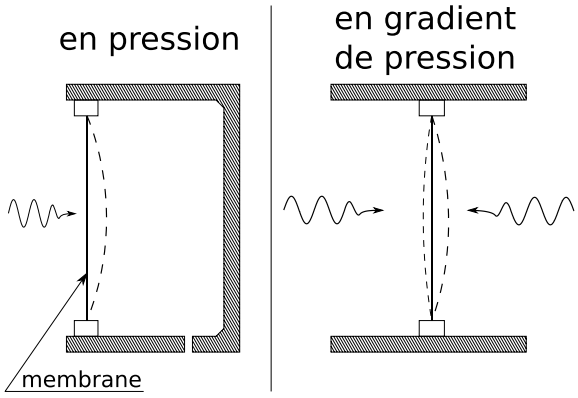

Figure 96. Structure d'un microphone « en pression » et « en gradient de pression ».

Le produit de la pression acoustique par la surface de la membrane donne la force imposée à celle-ci (Figure 96, gauche). On considère ici un des types principaux de microphone nommé capteur de « pression ». Il y a un autre type nommé microphone à « gradient de pression » pour lequel les ondes acoustiques sont en contact avec les deux faces de la membrane (Figure 96, droite). Dans ce cas, pour obtenir la force totale imposée sur la membrane il faut prendre en compte la différence des pressions acoustiques entre les deux faces de la membrane.

Le capteur acoustique dit « de pression » capte le son dans toutes les directions. La représentation du champ sonore se fait selon une directivité « omnidirectionnelle » (Figure 97) : toutes les directions d'incidence sont captées avec la même amplitude.

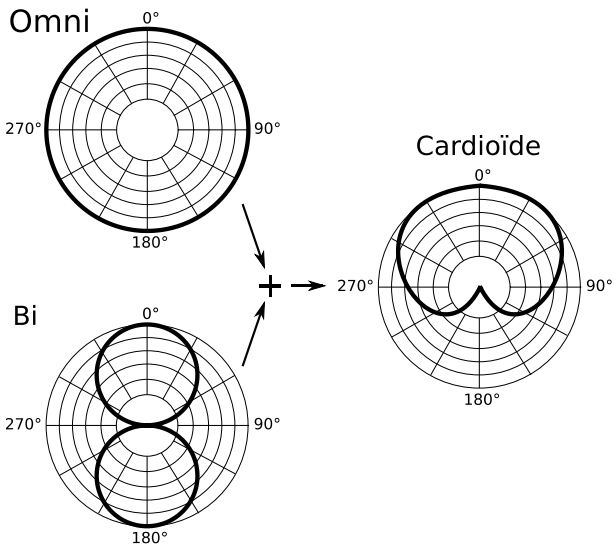

Figure 97. Directivité de microphone : Omnidirectionnelle, Bidirectionnelle et Cardioïde.

En pratique, les capteurs acoustiques sont intégrés dans le corps des microphones qui créent un obstacle pour les ondes acoustiques incidentes. Donc l'amplification ou l'atténuation des ondes dépendent des caractéristiques du corps de microphone et ne sont pas les même selon les bandes de fréquence.

Les ondes acoustiques qui arrivent vers la membrane d'un capteur dit à « gradient de pression » forment une directivité « en huit » ou bidirectionnelle (Figure 97). Les ondes acoustiques qui proviennent perpendiculairement par rapport à la membrane ne provoquent aucune vibration mécanique et par conséquence aucune tension dans le circuit du microphone.

Ces deux types de capteurs acoustiques permettent de combiner le principe de « pression » et de « gradient de pression », et réaliser d'autres diagrammes intermédiaires grâce à une combinaison linéaire de ceux-ci : $(\lambda \cdot \text{Omni} + \mu \cdot \text{Bi})$. Par exemple pour obtenir une figure de directivité cardioïde il suffit de faire la sommation sous la forme $(0.5 \cdot \text{Omni} + 0.5 \cdot \text{Bi})$ (Figure 97). Si λ est inférieur à μ, les formes typiques des diagrammes sont représentées sur la Figure 98.

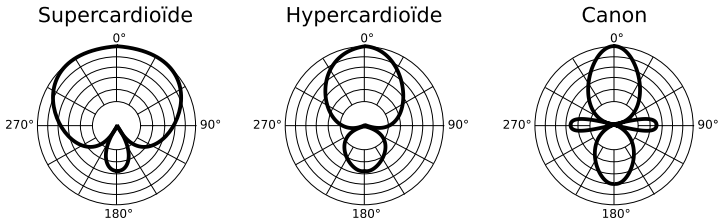

Figure 98. Directivités intermédiaires d'un microphone.

Techniques de captation stéréo

Technique XY : c'est la superposition de deux microphones avec une directivité de type « cardioïde » (pour autant que leurs corps le permettent) (Figure 99). L'angle de 90° entre les deux microphones est le plus utilisé. Plus l'angle est large et plus le champ stéréo perçu sera étendu [4]. Le « cardioïde » capte bien le champ sonore avant et permet d'éviter le champ arrière (par exemple le public pendant un concert).

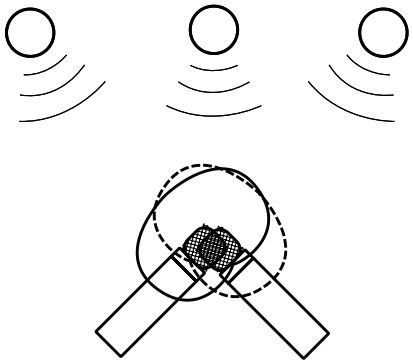

Figure 99. Captation du son par le microphone « stéréo ». Technique XY.

Technique Blumlein : la différence principale par rapport à la technique XY est l'utilisation de deux microphones bidirectifs (Figure 100) qui sont orientés de façon à former un angle de 90 degrés, leurs côtés positifs faisant face aux côtés gauche et droit de la source sonore [4]. Le diagramme « en huit » dans cette technique donne une bonne séparation stéréo

et en même temps, il permet de capter la réponse de la salle et éviter les ondes sonores latérales par rapport au micro. La technique Blumlein présente des inconvénients qui sont liés aux réflexions dans la salle. L'arrière du microphone gauche (Figure 100) capte également le son réfléchi de la partie droite et en même temps le lobe positif peut capter des ondes acoustiques réfléchies de l'avant, ce qui détériore la localisation [4].

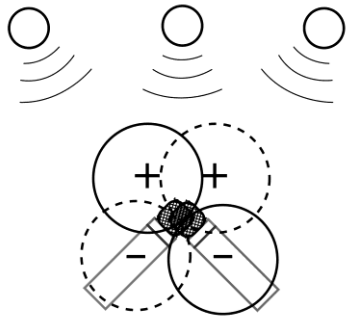

Figure 100. Captation du son par le microphone « stéréo ». Technique Blumlein.

Technique ORTF : cette technique a été créée par l'Office de Radio Télévision Française. Deux microphones sont placés à la distance de 17 cm l'un de l'autre et forment un angle de 110 degrés (Figure 101). L'idée principale est d'imiter la position des oreilles sur la tête humaine par deux microphones cardioïdes qui donnent la même profondeur que la technique Blumlein mais en même temps ne captent pas des réflexions à l'arrière grâce à la directivité des microphones.

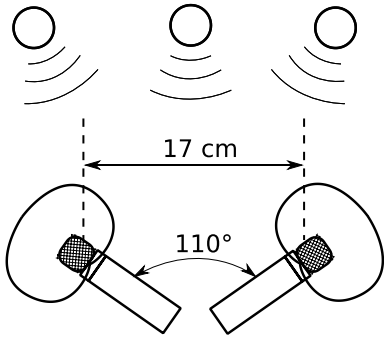

Figure 101. Captation du son par le microphone « stéréo ». Technique ORTF.

Technique MS (Mid-Side) : un microphone (Mid) de type « cardioïde » est placé en face de la source acoustique. Un deuxième microphone (Side) « en huit » prend sa position latérale par rapport au « Mid » (Figure 102). Le microphone « cardioïde » dans cette configuration capte bien la source acoustique et évite le bruit ou les effets sonores à l'arrière tandis que le microphone « en huit » reçoit des réflexions et obtient l'information sonore latérale. Cette configuration permet aux ingénieurs du son de régler séparément chaque microphone. Selon la scène sonore, le microphone « cardioïde » peut être remplacé par un omnidirectionnel pour

pouvoir capter le public. Cette technique utilise une méthode d'encodage matriciel à partir de signaux enregistrés pour obtenir la stéréo, ce qui est un premier pas vers l'encodage HOA.

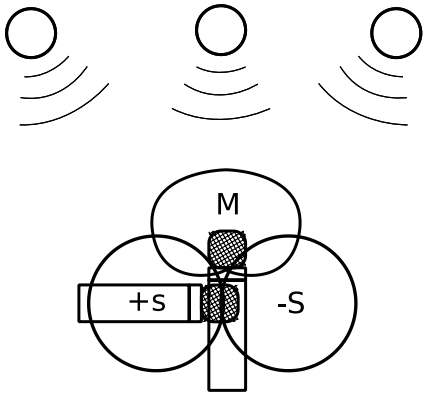

Figure 102. Captation du son par le microphone « stéréo ». Technique MS.

Technique AB (Omni écartés) : deux microphones omnidirectionnels sont écartés et souvent positionnés entre 1m20 et 2m50 face à la source acoustique [4] et à la même hauteur que les musiciens (Figure 103). Les Omni écartés donnent une bonne image sonore avec de la profondeur, par contre le champ sonore au centre est moins net. De nombreux ingénieurs du son préfèrent utiliser la technique AB pour enregistrer une ambiance extérieure ou de grandes orgues [4].

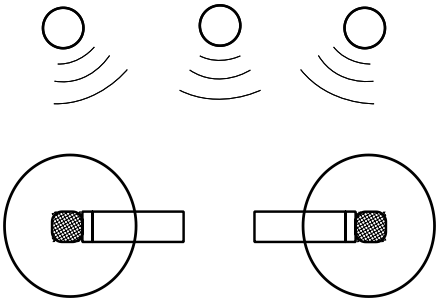

Figure 103. Captation du son par le microphone « stéréo ». Technique AB.

Technique de l'arbre Decca : les ingénieurs de Decca Records ont proposé l'arbre Decca (ou « Decca Tree ») qui est souvent utilisé dans l'enregistrement de musiques de films [4]. Trois microphones omnidirectionnels se trouvent à trois extrémités d'une structure de forme en « T » (Figure 104). Cette structure est habituellement montée à environ 2m50 – 3m50 du sol de façon à ce que le micro central se trouve juste derrière la tête du chef d'orchestre. La technique de l'arbre Decca grâce à deux microphones sur l'axe horizontal donne une bonne image des réflexions de la pièce tandis que le micro médian offre une bonne qualité du son central.

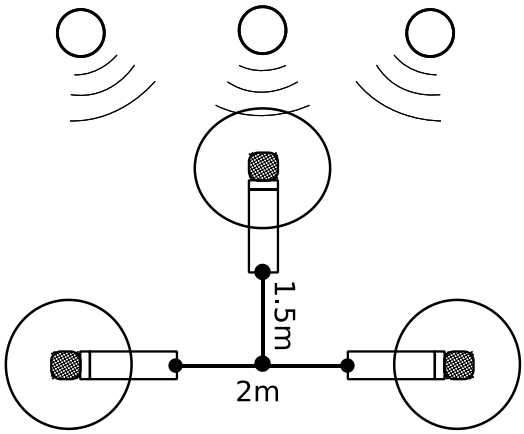

Figure 104. Captation du son par le microphone « stéréo ». Technique de l'arbre Decca.

Mixage stéréo

Il existe plusieurs approches, nommées lois de pan-pot [6], [31]. On considère les gains de chaque haut-parleur gauche et droite G_L et G_R, l'angle φ qui sépare chaque haut-parleur de l'axe médian et l'angle θ décrivant la position de la source acoustique (Figure 105).

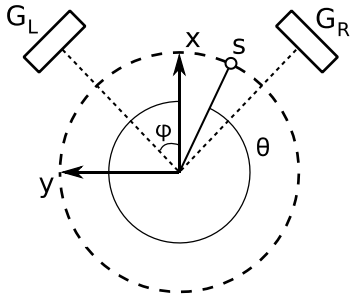

Figure 105. Répartition du son entre deux haut-parleurs.

A partir de la liberté de mouvement accordée à la tête, on exprime deux lois de pan-pot de la forme suivante [6], [31]:

$$\begin{cases} \dfrac{\sin\theta}{\sin\varphi} = \dfrac{G_L - G_R}{G_L + G_R} \\ \dfrac{G_R}{G_L} = \dfrac{\sin\varphi - \sin\theta}{\sin\varphi + \sin\theta} \end{cases} \tag{104}$$

$$\begin{cases} \dfrac{\tan\theta}{\tan\varphi} = \dfrac{G_L - G_R}{G_L + G_R} \\ \dfrac{G_R}{G_L} = \dfrac{\tan\varphi - \tan\theta}{\tan\varphi + \tan\theta} \end{cases} \tag{105}$$

La première équation (loi de sinus) dans le système

(104) et (105) décrit la position de la source (θ) pour une tête fixe dirigée vers l'axe médian. La deuxième (loi des tangentes) est dédiée à une tête libre ou dirigée vers la source virtuelle.

Pour prédire l'effet subjectif de localisation d'une source sonore on va utiliser le modèle de prédiction à l'aide d'un vecteur vélocité $\vec{V}$ dans le domaine de basses fréquences [32] et un vecteur d'énergie $\vec{E}$ dans le domaine des moyennes et hautes fréquences [33].

En introduisant un vecteur vélocité $\vec{V}$ [33] qui indique la direction d'incidence $\vec{u}_V$ du front d'onde reproduit [6] on peut réécrire le système d'équations

(104) et (105):

$$\vec{V} = \frac{G_L\vec{u}_L + G_R\vec{u}_R}{G_L + G_R} = \cos\varphi\,\vec{u}_x + \frac{G_L - G_R}{G_L + G_R}\sin\varphi\,\vec{u}_y \qquad (106)$$

où la composante suivant l'axe médian $\vec{u}_x$ est fixe et ne dépend que de l'angle φ [6].

Par analogie, dans le domaine des hautes fréquences, on introduit le vecteur énergie $\vec{E}$ [33] qui s'écrit comme une pondération des incidences $\vec{u}_L$ et $\vec{u}_R$ des haut-parleurs par les puissances associées G_L^2 et G_R^2 :

$$\vec{E} = \frac{G_L^2\vec{u}_L + G_R^2\vec{u}_R}{G_L^2 + G_R^2} = \cos\varphi\,\vec{u}_x + \frac{G_L^2 - G_R^2}{G_L^2 + G_R^2}\sin\varphi\,\vec{u}_y \qquad (107)$$

En formulation vectorielle l'ensemble des équations

(104)-(107) peut être présenté sous la forme [34] (VBAP – Vector Base Amplitude Panning):

$$\vec{u}_S = g_1\vec{u}_1 + g_2\vec{u}_2 \qquad (108)$$

où $\vec{u}_1$ et $\vec{u}_2$ sont les directions des deux haut-parleurs, $\vec{u}_S$ la direction d'une source restituée avec l'angle θ. Le vecteur $\vec{u}_S$ correspond au vecteur vélocité. Au final deux méthodes de prédiction d'une source acoustique dans les basses et hautes fréquences par le vecteur de vélocité et énergie s'expriment :

$$\vec{V} = \frac{g_1\vec{u}_1 + g_2\vec{u}_2}{g_1 + g_2}$$
$$\vec{E} = \frac{\sum g_i^2\vec{u}_i}{\sum g_i^2} \tag{109}$$

A partir de l'équation (108) on déduit la forme matricielle :

$$\boldsymbol{u_S^t} = \boldsymbol{g U_{12}}$$
$$\boldsymbol{g} = [g_1\ g_2]^t$$
$$\boldsymbol{U_{12}} = [\boldsymbol{u_1}\ \boldsymbol{u_2}]^t = [\vec{u}_1\ \vec{u}_2]^t = \begin{bmatrix}\cos\theta_1 & \sin\theta_1\\ \cos\theta_2 & \sin\theta_2\end{bmatrix} \tag{110}$$

Annexe 2 : Higher Order Ambisonics

HOA Extension vers les ordres supérieurs

L'ordre 1 Ambisonic décrit le champ sonore indépendamment de système des haut-parleurs tandis que High Order Ambisonics compte tenu des avantages de l'ordre 1 permet d'améliorer la résolution spatiale et élargir la zone d'écoute [6], [35]. Pour passer aux ordres supérieurs Jérôme Daniel a proposé la solution qui consiste en la décomposition du champ sonore par des harmoniques sphériques [36]. Les harmoniques sphériques en fonctions réelles s'expriment sous la forme :

$$Y_{mn}^{\sigma}(\theta,\delta) = \underbrace{\sqrt{(2m+1)\varepsilon_n \frac{(m-n)!}{(m+n)!}}}_{normalisation} \cdot \underbrace{P_{mn}(\sin\theta)}_{\substack{polynômes \\ de\ Legendre\ associés}} \times \begin{cases} \cos n\theta\ si\ \sigma = 1 \\ \sin n\theta\ si\ \sigma = -1 \end{cases} \quad (111)$$

où θ et δ sont respectivement l'azimut et l'élévation, m et n sont des entiers positifs tels que $n \leq m$, σ prend les valeurs +1 et -1, et ε_n est égal à 1 si $n = 0$ et égal à 2 si $n > 2$. Les polynômes de Legendre associés sont définis pour tout x appartenant à $[-1, 1]$ par :

$$P_{mn}(x) = (1-x^2)^{\frac{n}{2}} \frac{d^n}{dx^n} P_m(x) \quad (112)$$

avec $m \geq 0$, et où $P_m(x)$ est le polynôme de Legendre de première espèce d'ordre m, qui peut être calculé numériquement grâce à une relation de récurrence :

$$\begin{cases} P_0(x) = 1 \\ P_1(x) = x \\ (m+1)P_{m+1}(x) = (2m+1)xP_m(x) - mP_{m-1}(x), \ \ m > 1 \end{cases} \quad (113)$$

Les harmoniques sphériques et leurs composantes pour les ordres jusqu'à quatre sont disponibles dans le Tableau 3.

Ordre m	Fonction $Y_{mn}^{\sigma}(\theta,\delta)$	Normalisation $\sqrt{(2m+1)\varepsilon_n \frac{(m-n)!}{(m+n)!}}$	Legendre associés $P_{mn}(\sin\delta)$	$\begin{cases}\cos n\theta\ si\ \sigma = 1 \\ \sin n\theta\ si\ \sigma = -1\end{cases}$	Composante
0	$Y_{00}^{1}(\theta,\delta)$	1	1	1	W
1	$Y_{11}^{1}(\theta,\delta)$	$\sqrt{3}$	$\cos\delta$	$\cos\theta$	X
	$Y_{11}^{-1}(\theta,\delta)$	$\sqrt{3}$	$\cos\delta$	$\sin\theta$	Y

	$Y_{10}^{1}(\theta,\delta)$	$\sqrt{3}$	$\sin\delta$	1	Z
2	$Y_{22}^{1}(\theta,\delta)$	$\sqrt{5/12}$	$3\cos^2\delta$	$\cos 2\theta$	U
	$Y_{22}^{-1}(\theta,\delta)$	$\sqrt{5/12}$	$3\cos^2\delta$	$\sin 2\theta$	V
	$Y_{21}^{1}(\theta,\delta)$	$\sqrt{5/3}$	$3\cos\delta\sin\theta$	$\cos\theta$	S
	$Y_{21}^{-1}(\theta,\delta)$	$\sqrt{5/3}$	$3\cos\delta\sin\theta$	$\sin\theta$	T
	$Y_{20}^{1}(\theta,\delta)$	$\sqrt{5}$	$(3\sin^2\delta-1)/2$	1	R
3	$Y_{33}^{1}(\theta,\delta)$	$\sqrt{7/360}$	$15\cos^3\delta$	$\cos 3\theta$	
	$Y_{33}^{-1}(\theta,\delta)$	$\sqrt{7/360}$	$15\cos^3\delta$	$\sin 3\theta$	
	$Y_{32}^{1}(\theta,\delta)$	$\sqrt{7/60}$	$15\cos^2\delta\sin\delta$	$\cos 2\theta$	
	$Y_{32}^{-1}(\theta,\delta)$	$\sqrt{7/60}$	$15\cos^2\delta\sin\delta$	$\sin 2\theta$	
	$Y_{31}^{1}(\theta,\delta)$	$\sqrt{7/6}$	$3\cos\delta\,(5\sin^2\delta-1)/2$	$\cos\theta$	
	$Y_{31}^{-1}(\theta,\delta)$	$\sqrt{7/6}$	$3\cos\delta\,(5\sin^2\delta-1)/2$	$\sin\theta$	
	$Y_{30}^{1}(\theta,\delta)$	$\sqrt{5}$	$\sin\delta\cdot(5\sin^2\delta-3)/2$	1	
4	$Y_{44}^{1}(\theta,\delta)$	$\sqrt{1/2240}$	$105\cos^4\delta$	$\cos 4\theta$	
	$Y_{44}^{-1}(\theta,\delta)$	$\sqrt{1/2240}$	$105\cos^4\delta$	$\sin 4\theta$	
	$Y_{43}^{1}(\theta,\delta)$	$\sqrt{1/280}$	$105\cos^3\delta\sin\delta$	$\cos 3\theta$	
	$Y_{43}^{-1}(\theta,\delta)$	$\sqrt{1/280}$	$105\cos^3\delta\sin\delta$	$\sin 3\theta$	
	$Y_{42}^{1}(\theta,\delta)$	$\sqrt{1/20}$	$15\cos^2\delta\,(7\sin^2\delta-1)/2$	$\cos 2\theta$	
	$Y_{42}^{-1}(\theta,\delta)$	$\sqrt{1/20}$	$15\cos^2\delta\,(7\sin^2\delta-1)/2$	$\sin 2\theta$	
	$Y_{41}^{1}(\theta,\delta)$	$\sqrt{9/10}$	$5\sin\delta\cos\delta\,(7\sin^2\delta-3)/2$	$\cos\theta$	
	$Y_{41}^{-1}(\theta,\delta)$	$\sqrt{9/10}$	$5\sin\delta\cos\delta\,(7\sin^2\delta-3)/2$	$\sin\theta$	
	$Y_{40}^{1}(\theta,\delta)$	$\sqrt{9}$	$(35\sin^4\delta-30\sin^2\delta+3)/8$	1	

Tableau 3. Les composantes des harmoniques sphériques pour l'ordre 0 à 4.

En décrivant le champ acoustique en coordonnées sphériques à l'aide de l'équation d'Helmholtz homogène [6] c'est possible de l'exprimer sous la forme de la série de Fourier-Bessel sphérique [37] :

$$p(r,\theta,\delta)=\sum_{m=0}^{+\infty} i^m j_m(kr)\sum_{n=0}^{m}\sum_{\sigma=\pm 1} B_{mn}^{\sigma}Y_{mn}^{\sigma}(\theta,\delta) \tag{114}$$

où $p(r,\theta,\delta)$ est la pression acoustique, B_{mn}^{σ} sont des coefficients de Fourier sphérique, $Y_{mn}^{\sigma}(\theta,\delta)$ sont les harmoniques sphériques, $j_m(kr)$ sont des fonctions de Bessel sphériques [6], [35] avec le nombre d'onde $(k = 2\pi f/c)$ qui s'expriment comme :

$$\begin{cases} j_0(0) = 1 \\ j_m(x) = (-1)^m x^m \left(\frac{1}{x}\frac{d}{dx}\right)^m \frac{\sin x}{x} \end{cases} \quad (115)$$

L'encodage ou la décomposition du champ sonore à un ordre M (autrement dit l'approximation) peut être décrit sous la forme suivante par application de l'équation (114) :

$$p_M(r,\theta,\delta) = \sum_{m=0}^{M} i^m j_m(kr) \sum_{n=0}^{m} \sum_{\sigma=\pm 1} B_{mn}^{\sigma} Y_{mn}^{\sigma}(\theta,\delta) \quad (116)$$

Les signaux B_{mn}^{σ} $(0 \le m \le M)$ définissent une représentation ambisonique 3D, homogène, d'ordre M [6] et en même temps développent le format B, c'est-à-dire :

$$\begin{cases} B_{00}^{1} = W \\ B_{11}^{1} = X \\ B_{11}^{-1} = Y \\ B_{10}^{1} = Z \\ \dots \end{cases} \quad (117)$$

Chaque ordre $[0 \dots M]$ comporte $(2m+1)$ coefficients. Le nombre K total (canaux ambisoniques) se calcule par l'équation :

$$\begin{aligned} K^{2D} &= 2M + 1 \\ K^{3D} &= \sum_{m=0}^{M} (2m+1) = (M+1)^2 \end{aligned} \quad (118)$$

Le nombre de composantes HOA est présenté dans le Tableau 4:

	Ordre	Nombre de signaux HOA
2D	0	1
	1	3
	2	5
	3	7
	m	$2m+1$
3D	0	1
	1	4
	2	9

	3	16
	4	25
	m	$(m+1)^2$

Tableau 4. Nombre de signaux HOA.

Les équations d'encodage HOA d'une onde plane d'incidence $s(t)$ s'exprime sous la forme [6]:

$$B_{mn}^{\sigma} = S(\omega) Y_{mn}^{\sigma}(\theta, \delta) \tag{119}$$

avec $0 \leq n \leq m$, $\sigma = \pm 1$

L'équation (119) à l'ordre M peut s'écrire sous une forme matricielle :

$$\mathbf{b}_M^{3D} = \mathbf{g}_M^{3D} \cdot S(\omega) \tag{120}$$

avec

$$\mathbf{b}_M^{3D} = \left(\underbrace{B_{00}^1 \, B_{11}^1 \, B_{11}^{-1} \, B_{10}^0}_{W,X,Y,Z} \ldots \underbrace{B_{mm}^1 B_{mm}^{-1} \ldots B_{mn}^1 B_{mn}^{-1} \ldots B_{m0}^1}_{2m+1} \ldots \underbrace{B_{MM}^1 B_{MM}^{-1} \ldots B_{M0}^1}_{2M+1} \right)^t \tag{121}$$

et

$$\mathbf{g}_M^{3D} = \left(Y_{00}^1 \, Y_{11}^1 \, Y_{11}^{-1} \, Y_{10}^0 \ldots \underbrace{Y_{mm}^1 Y_{mm}^{-1} \ldots Y_{mn}^1 Y_{mn}^{-1} \ldots Y_{m0}^1}_{2m+1} \ldots \underbrace{Y_{MM}^1 Y_{MM}^{-1} \ldots B_{M0}^1}_{2M+1} \right)^t \tag{122}$$

Le vecteur $\mathbf{b}_M^{3D}$ décrit les signaux HOA à l'ordre Met le vecteur $\mathbf{g}_M^{3D}$ les gains d'encodage associés [6].

Décodage HOA

L'étape de décodage spatial ou de reconstruction de certaines caractéristiques du champ sonore consiste en l'application d'une matrice de gains aux signaux HOA [6] qui permet de restituer le champ sonore sur une sphère de haut-parleurs (3D) ou sur un cercle (2D). On considère que les ondes sonores émises par les haut-parleurs sont planes dans la zone d'écoute. En trois dimensions le champ reconstruit s'exprime :

$$\mathbf{C} \cdot \mathbf{s} = \mathbf{b} \tag{123}$$

avec

$$\mathbf{C} = \begin{bmatrix} Y_{00}^{1}(\theta_1, \delta_1) & Y_{00}^{1}(\theta_2, \delta_2) & \cdots & Y_{00}^{1}(\theta_L, \delta_L) \\ \vdots & \vdots & \vdots & \vdots \\ Y_{11}^{1}(\theta_1, \delta_1) & Y_{11}^{1}(\theta_2, \delta_2) & \cdots & Y_{11}^{1}(\theta_L, \delta_L) \\ \vdots & \vdots & \vdots & \vdots \\ Y_{M0}^{1}(\theta_1, \delta_1) & Y_{M0}^{1}(\theta_2, \delta_2) & \cdots & Y_{M0}^{1}(\theta_L, \delta_L) \end{bmatrix},$$

$$\mathbf{s} = \begin{bmatrix} S_1 \\ S_2 \\ \vdots \\ S_L \end{bmatrix}, \mathbf{b} = \begin{bmatrix} B_{00}^{1} \\ \vdots \\ B_{mn}^{\sigma} \\ \vdots \\ B_{M0}^{1} \end{bmatrix} \tag{124}$$

où L est le nombre de haut-parleurs, la matrice $\mathbf{C}$ contient les vecteurs harmoniques sphériques associés à chaque direction de haut-parleur, le vecteur $\mathbf{s}$ contient les signaux émis par les L haut-parleurs et le vecteur $\mathbf{b}$ contient les signaux HOA.

La solution de l'équation (124) dépend des propriétés de la matrice $\mathbf{C}$ (elle doit être inversible). La solution exacte est déduite à partir de la matrice notée $\mathbf{D}$ (la matrice de décodage ou pseudo-inverse de Moore-Penrose [37]) :

$$\mathbf{s} = \mathbf{D} \cdot \mathbf{b} \text{ avec } \mathbf{D} = \mathbf{C^t} \cdot \ (\mathbf{C} \cdot \mathbf{C^t})^{-1} \tag{125}$$

Pour restituer les signaux HOA décodés vers un casque d'écoute il faut passer par une autre étape complémentaire qui consiste à effectuer la *binauralisation*. Le terme « binaural » (*bin* – deux, *aural* - oreilles) est un autre moyen de présenter le son restitué (soit enregistré soit traité) en prenant en compte des paramètres du système auditif humain tels que la forme des pavillons, la dimension du corps et de la tête qui jouent un rôle important dans la réflexion et la transmission du son. Parmi tous les paramètres utiles pour localiser la source acoustique dans l'espace, on utilise certains indices comme la différence de temps, d'intensité et les indices spectraux [38]. Les indices peuvent former une base de réponses impulsionnelles ou des fonctions de transfert (dans le domaine fréquentiel) nommées HRTF (*Head-Related Transfer Functions*) permettant d'extraire la position angulaire de la source acoustique pour chaque tympan. Le moyen de binauraliser le son est de passer les signaux HOA à travers un filtre contenant l'information des HRTF. En exprimant les signaux binauraux pour l'oreille gauche et droite comme S_G et S_D on peut les déduire à partir d'une paire d'HRTF $H_G(\theta_l, \delta_l)$ et $H_D(\theta_l, \delta_l)$:

$$\begin{cases} S_G = \sum_{l=1}^{L} H_G(\theta_l, \delta_l) \cdot S_l = \mathbf{h}_G^t \cdot \mathbf{s} \\ S_D = \sum_{l=1}^{L} H_D(\theta_l, \delta_l) \cdot S_l = \mathbf{h}_D^t \cdot \mathbf{s} \end{cases} \quad (126)$$

avec le haut-parleur l qui émet le son $\mathbf{s}$ (cf. l'équation (124)). $\mathbf{h}_G^t$ et $\mathbf{h}_D^t$ contiennent respectivement les HRTF :

$$\begin{cases} \mathbf{h}_G^t = (H_G(\theta_1, \delta_1) \;\; H_G(\theta_2, \delta_2) \;\; \cdots \;\; H_G(\theta_l, \delta_l)) \\ \mathbf{h}_D^t = (H_D(\theta_1, \delta_1) \;\; H_D(\theta_2, \delta_2) \;\; \cdots \;\; H_D(\theta_l, \delta_l)) \end{cases} \quad (127)$$

En utilisant l'équation (125) S_G et S_D s'exprime sous la forme suivante :

$$\begin{cases} S_G = \mathbf{h}_G^t \cdot \mathbf{D} \cdot \mathbf{b} \\ S_D = \mathbf{h}_D^t \cdot \mathbf{D} \cdot \mathbf{b} \end{cases} \quad (128)$$

Annexe 3 : Projection cartographique

La projection cartographique est une représentation géométrique d'une image sphérique sur un plan azimutal. Cette approche permet de garder toute l'information angulaire et spatiale [39]. Dans ce travail on considère une projection de Mercator [39] (Figure 106) qui fait la partie de projection cylindrique.

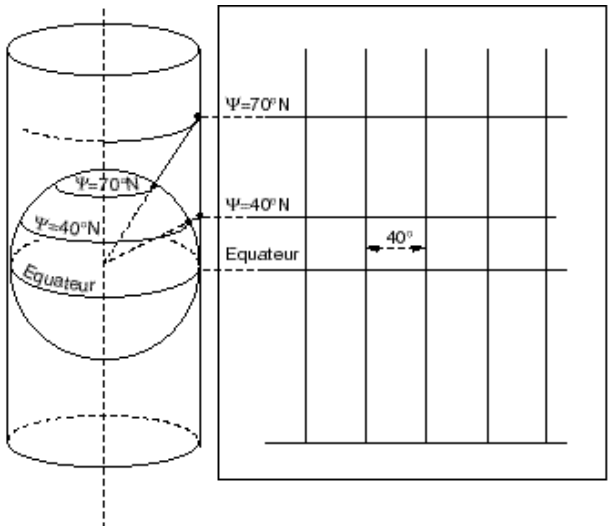

Figure 106. Projection de Mercator

A la base la projection cylindrique ne conserve pas les angles. Il est facile de montrer cette propriété basée sur la transformation en plan horizontal (x, y) suivante :

$$x = \cos(\varphi_0)(\theta - \theta_0) \qquad (129)$$
$$y = (\varphi - \varphi_0)$$

où φ_0 et θ_0 sont constantes. Dans le cas particulier ($\varphi_0 = 0$ et $\theta_0 = 0$) on obtient :

$$x = \theta \qquad (130)$$
$$y = \varphi$$

Pour conserver l'information spatiale Mercator [39] a modifié la transformée entre les coordonnées polaires et cartésiennes par les équations suivantes :

$$x = (\theta - \theta_0) \qquad (131)$$
$$y = \ln\left[\tan\left(\frac{\pi}{4} + \frac{\varphi}{2}\right)\right] = \frac{1}{2}\ln\left(\frac{1+\sin\varphi}{1-\sin\varphi}\right) = \sinh^{-1}(\tan\varphi)$$

Pour passer auxcordonnées polaires il faut effectuer une inversion basée sur la fonction de Gudermann inverse [39] :

$$\theta = 2\tan^{-1}(e^{y}) - \frac{1}{2}\pi = \tan^{-1}(\sinh y) \tag{132}$$
$$\varphi = x + \theta_0$$

Bibliographie

[1] Andrey Fedosov, Grégory Pallone, Jérôme Daniel, Sylvain Marchand, “Automatic HOA Mixing,” in *International Conference on Spatial Audio (ICSA)*, Graz, Austria, 2015.

[2] Andrey Fedosov, Jérôme Daniel, Gregory Pallone, “Procédé de traitement de données pour l’estimation de paramètres de mixage de signaux audio, procédé de mixage, dispositifs, et programmes d’ordinateurs associés,” Brevet FR3034892, 14-Oct-2016.

[3] David Marston, “The Audio Definition Model,” *Fourth W3C Web TV Workshop*, 2014.

[4] Aspen Pittman, “Prise de son stéréo,” 20-Apr-2003. [Online]. Available: http://fr.audiofanzine.com/microphone/editorial/dossiers/prise-de-son-stereo.html.

[5] Victoria Turk, “Alan Blumlein - the man who invented stereo,” 01-Apr-2015. [Online]. Available: http://www.abbeyroad.com/News/Article/22/Alan-Blumlein-the-man-who-invented-stereo.

[6] J. Daniel, “Représentation de champs acoustiques, application a la transmission et a la restitution de scènes sonores complexes dans un contexte multimedia,” Thèse, Paris 6, 2000.

[7] Pierre-Antoine Signoret, “Le son multicanal.” 2012.

[8] M. A. Gerzon, “Periphony: With-Height Sound Reproduction,” *J. Audio Eng. Soc.*, vol. 21, no. 1, pp. 2–10, Feb. 1973.

[9] D. Mercier, *Le livre des techniques du son - 4e édition: Tome 3 - L’exploitation*. Dunod, 2013.

[10] P. G. Craven and M. A. Gerzon, “Coincident microphone simulation covering three dimensional space and yielding various directional outputs,” Brevet US 4042779, 16-Aug-1977.

[11] mhAcoustics, “Eigenmike HOA microphone,” 2009. [Online]. Available: http://www.mhacoustics.com/.

[12] Matthias Kronlachner, “Plug-in Suite for Mastering the Production and Playback in Surround Sound and Ambisonics,” presented at the 136th AES Convention Berlin, 2013.

[13] Blue Ripple Sound, “Third Order Ambisonics (TOA plugins).” [Online]. Available: http://www.blueripplesound.com/.

[14] Centre de recherche Informatique et Création Musicale (CICM), “La Bibliothèque Hoa,” 2013. [Online]. Available: http://www.mshparisnord.fr/hoalibrary/telechargements/.

[15] Penha, R., & Oliveira, J.P, “Spatium, tools for sound spatialization,” presented at the Proceedings of the Sound and Music Computing Conference, 2013.

[16] TwoBigEars, “Spatial Workstation,” 2016. [Online]. Available: https://facebook360.fb.com/spatial-workstation/.

[17] A. K. Tellakula, “Acoustic Source Localization Using Time Delay Estimation,” Thèse, Supercomputer Education And Research Centre Indian Institute of Science, 2009.

[18] J. J. Weng and K. Y. Guentchev, “Learning-Based Three Dimensional Sound Localization Using a Compact NonCoplanar Array of Microphones,” *AAAI*, 1998.

[19] P. Pertilä, *Acoustic Source Localization in a Room Environment and at Moderate Distances*, vol. 794. Tampere University of Technology, 2009.

[20] J. Chen, J. Benesty, and Y. (Arden) Huang, “Time Delay Estimation in Room Acoustic Environments: An Overview,” *EURASIP J. Adv. Signal Process.*, vol. 2006, pp. 1–20, 2006.

[21] E. Gallo, N. Tsingos, and G. Lemaitre, “3D-Audio Matting, Post-editing and Re-rendering from Field Recordings,” *EURASIP J. Adv. Signal Process.*, p. 16, 2007.

[22] D. Barchiesi and J. Reiss, “Reverse engineering of a mix,” *J. Audio Eng. Soc.*, pp. 563–576, 2010.

[23] V. Pulkki, M.-V. Laitinen, J. Vilkamo, J. Ahonen, T. Lokki, and T. Pihlajamäki, “Directional audio coding - perception-based reproduction of spatial sound,” International Workshop on the Principles and Applications of Spatial Hearing, 2009, pp. 1–4.

[24] J. Vilkamo, “Spatial Sound Reproduction with Frequency Band Processing of B-Format Audio Signals,” Thèse, Helsinki University of Technology, 2016.

[25] H. C. So, “On time delay estimation using an FIR filter,” *Signal Process.*, vol. 81, no. 8, pp. 1777–1782, Aug. 2001.

[26] D. Ventzas and N. Petrellis, “Peak Searching Algorithms and Applications,” *Int. Conf. Signal Image Process. Appl. - SIPA*, 2011.

[27] Sven Vörtmann, “VST tecnology, 3rd Party Developer,” 1996. [Online]. Available: http://www.steinberg.net/en/company/developers.html.

[28] K. W. Wilson, S. Member, and T. Darrell, “Learning a precedence effect-like weighting function for the generalized cross-correlation framework,” *IEEE J. Of*, 2006.

[29] J. Choi, J. Kim, and N. S. Kim, “Robust Time-Delay Estimation for Acoustic Indoor Localization in Reverberant Environments,” *IEEE Signal Process. Lett.*, vol. 24, no. 2, pp. 226–230, Feb. 2017.

[30] S. Busson, “Individualization of acoustic cues for binaural synthesis, Individualization of acoustic cues for binaural synthesis,” Thèse, Université de la Méditerranée - Aix-Marseille II, Université de la Méditerranée - Aix-Marseille II, 2006.

[31] B. Bernfeld, “Simple Equations for Multichannel Stereophonic Sound Localization,” *J. Audio Eng. Soc.*, vol. 23, no. 7, pp. 553–557, Sep. 1975.

[32] Y. Makita, “On the Directional Localization of Sound,” *EBU Rev.*, vol. 73, p. 1536:1539, 1962.

[33] M. A. Gerzon, “General Metatheory of Auditory Localisation,” presented at the Audio Engineering Society Convention 92, 1992.

[34] V. Pulkki, “Virtual Sound Source Positioning Using Vector Base Amplitude Panning,” *J. Audio Eng. Soc.*, vol. 45, no. 6, pp. 456–466, Jun. 1997.

[35] S. Moreau, “Étude et réalisation d’outils avancés d’encodage spatial pour la technique de spatialisation sonore Higher Order Ambisonics: microphone 3D et contrôle de distance,” Thèse, Université du Maine, France, 2006.

[36] W. Appel, *Mathématiques pour la physique et les physiciens !*, Édition : 4e édition. Paris: H&K, 2008.

[37] G. Dauphin, “Notes de cours. Traitement du signal.,” Université Paris Nord. Laboratoire de Traitement et Transport de L’Information (L2TI), 2013.

[38] J. Blauert, *Spatial Hearing - Revised Edition: The Psychophysics of Human Sound Localization*, Revised edition edition. Cambridge, Mass: The MIT Press, 1996.

[39] E. W. Weisstein, “Mercator Projection,” 1999. [Online]. Available: http://mathworld.wolfram.com/MercatorProjection.html.

{Paper from ICSA2015 Proceedings, Graz, 3rd International Conference on Spatial Audio}

Automatic HOA mixing

Andrey Fedosov[1], Gregory Pallone[2], Jérôme Daniel[3], Sylvain Marchand[4]
[1] *b<>com, 35510, Cesson-Sévigné, France, Email: andrey.fedosov@b-com.com*
[2] *b<>com, 35510, Cesson-Sévigné, France, Email: gregory.pallone @b-com.com*
[3] *b<>com, 35510, Cesson-Sévigné, France, Email: jerome.daniel @b-com.com*
[4] *Lab-STICC – CNRS, University of Brest, 29238 Brest, France, Email: Sylvain.Marchand@univ-brest.fr*

Abstract

Thanks to the growing interest in 3D sound and the availability of 3D microphones, but also considering HOA (Higher Order Ambisonics) advantages (easiness of recording, independency of the sound pickup and rendering setup), and support of this format in the new MPEG-H 3D Audio codec, HOA recordings will probably increase in popularity during the next decade. In this context, we will study the purpose of sound engineers mixing HOA and spot microphones, and therefore the issue of estimating parameters such as delay, position and gain of acoustic sources associated to spot microphones. This approach of mixing main and spot microphones is widely used in classical music recordings, but it can also be used in different professional applications (theatre and cinema) as well as in mass-market applications (music rehearsal, family events recordings). We propose an algorithm providing estimated parameters (delay, position, gain) based on spatial encoding equations in the HOA format that would be used to process the spot microphone signals during the mix. The robustness of the estimators is evaluated on recorded and artificial sound scenes, with different degrees of complexity in terms of number of sources and acoustic conditions (reverberation, effect of real microphone encoding, source directivity…). This automatic parameters extraction can be seen as an assistance for sound engineers, avoiding them unattractive work such as measuring the distances and angles between microphones, and allowing them to concentrate on artistic issues such as adjusting levels, EQ or compressor parameters, and also fine adjustments of delays, positions and gains if necessary.

Introduction

A classical stereo sound recording relies primarily on the use of a stereo pair setup (referred to as "main microphone") providing a global sound image of the whole scene, while bringing space's color and volume. In order to enhance some of the acoustic sources, the sound engineer uses also one or several mono (or sometimes stereo) microphones (referred to as "spot microphones") located close to them. This allows a precise work on the sound of the sources in the mix, after having manually time-synchronized and panned the tracks.

This approach of mixing between main and spot microphones is widely used in classical music recordings, but it could also be used in different professional applications (theatrical and cinema sound recordings, where the spot microphones placed on actors can move, TV and radio live programs…) as well as in mass-market applications (music rehearsal, family events recordings…).

Nowadays, the rendering setups are evolving towards a higher number of loudspeakers (or binaural on headphone) to improve immersion: from stereo, the contents are now widely produced in surround (typically 5.1), and the next generation formats, called 3D, will allow to use elevation [1]. In this context, the Ambisonics technology proposed by Gerzon [2] is of great interest since it is based on a universal format naturally describing the 3D sound field and relying on a powerful mathematical formalism. This format, which allows simple transformations of the sound scene (rotation, focus), has the advantage of being independent from both the sound pickup and the speaker setup [2], [3]. This format, limited in terms of spatial resolution (first order of the spherical harmonics), has been generalized to higher orders in a so-called "Higher Order Ambisonics" (HOA) format [3]. Due to a combination of different factors such as growing interest in 3D sound, advantages of HOA (easiness of recording, independency of the sound pickup and rendering setup), availability of 3D microphones ([4], [5], [6], [7]), and support of HOA in the new MPEG-H 3D Audio codec [1], HOA recordings will probably increase in popularity during the next decade. In this context, it is essential to provide sound engineers with adapted tools allowing them to work with the same habits as they are used to with stereo and surround formats.

In this paper, we propose a global architecture allowing to mix a HOA "main" microphone with several mono "spot" microphones, in a framework adapted to most of the commercial audio editors (as long as their internal busses support the number of HOA components chosen for the project). We also provide a technical solution to extract automatically the parameters necessary for a correct mixing between microphones, which is evaluated on both artificial and real signals.

This automatic parameters extraction can be seen as an assistance for sound engineers, avoiding them unattractive work such as measuring the distances and angles between microphones, and allowing them to concentrate on artistic issues such as adjusting levels, EQ or compressor parameters, and also fine adjustments of delays, positions and gains if necessary.

Mixing issues

Figure 1, inspired by [8], shows a typical signal workflow used by a sound engineer mixing a classical music recording

with a main (HOA) microphone and N spot (mono) microphones, on any Digital Audio Workstation (DAW). The technology proposed in this paper is materialized in the "parameter extraction" module. This module can be implemented as a plugin in any DAW. Its interface takes as inputs the main and spot microphones signals, and as output the automatic parameters (delay, position, gain) needed for the processing modules for each of the spot microphones.

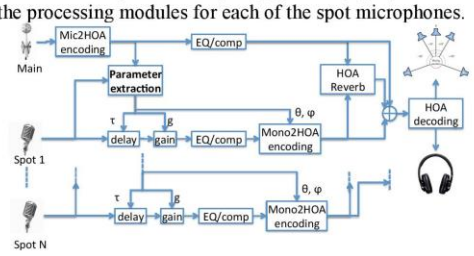

Figure 1: Workflow inside a DAW

The "delay" processing module allows to time-align the spot signal with the main signal in order to compensate for the propagation delay of the acoustic wave between microphones. An error on the delay value would lead to artifacts such as comb filter effect, echo or pre-echo.

The "gain" processing module allows adjusting the volume of the source inside the spot signal to the volume of the source inside the main signal.

The "Mono2HOA encoding" processing module aims at spatially encoding the spot microphone signal using the azimuth and elevation angles of the source as recorded by the main microphone. Details on related signal processing can be found in [9]. An error on the position value would lead to artifacts such as instability and/or enlargement of the source.

The "parameter extraction" module provides automatic parameter values, which remain to be adjusted artistically by the sound engineer. The HOA decoding for loudspeakers or headphones, and the use of the equalizer (EQ), compressor (comp), reverberation (HOA reverb) and summation in the HOA domain are not described in this paper but can be found in the literature [8] and in commercial products [10], [11].

General Formalism

Higher Order Ambisonics

Higher Order Ambisonics provides a spatial representation of the sound field that relies on its spherical harmonic decomposition, centered on a reference point that can be considered as the listener's viewpoint [3]. The ambisonic "order" is related to the truncation of the decomposition series and defines the spatial resolution. With regard to the analysis algorithm presented in this paper, we restrict the formalism to the first order ambisonics initially proposed by Gerzon [2]. It consists of the four components that would be captured at the reference point by: one omnidirectional microphone (providing pressure signal W) and three bidirectional microphones (components X, Y, Z) oriented along the orthogonal axes, where X (resp. Y, Z) usually points forwards (resp. left, upwards). This set of four components composes the so-called B-format, obtained from the A-format (signals of the capsules from microphones such as Soundfield ® or Eigenmike ®) by applying a matrix of gains or filters.

We consider a sound source emitting a signal s, and define its location as a point of the 3D Euclidean space described in terms of spherical coordinates (azimuth θ, elevation Φ, radius r) or Cartesian coordinates (x,y,z). The following conversion makes also appears the components of a unit vector $\vec{u}$ (u_x, u_y, u_z):

$$\begin{cases} x = r \cdot \cos\theta \cdot \cos\varphi = r \cdot u_x \\ y = r \cdot \sin\theta \cdot \cos\varphi = r \cdot u_y \\ z = r \cdot \sin\varphi = r \cdot u_z \end{cases} . \quad (1)$$

With the usual assumption that the source is in far field, the generated ambisonic sound field is described as follows [2] [3]:

$$\begin{cases} W = s \\ X = \eta \cdot s \cdot u_x \\ Y = \eta \cdot s \cdot u_y \\ Z = \eta \cdot s \cdot u_z \end{cases}, \quad (2)$$

where η is a normalization factor (e.g. $\eta = \sqrt{2}$ in [2]).

Sound scene capture

We consider a sound scene (**Figure 2**) composed of M acoustic sources emitting M signals $s_m(t)$, and captured by:

- N spot microphones yielding signals $a_n(t)$, each one as a result of a combination of the filtered sources,
- one main microphone that is typically an ambisonic or a HOA microphone array that yields a spatial description of the whole sound field through at least the 4 ambisonic components of the B-format [2], [3].

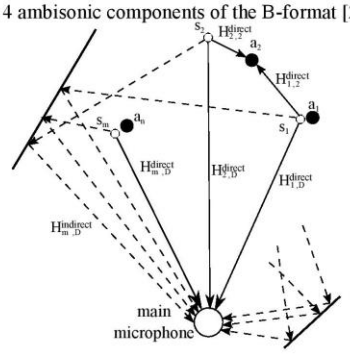

Figure 2: Sound scene composed of M acoustic sources.

All the captured signals can be expressed as:

$$D(t) = \sum_{m=1}^{M} \left\{ \left[h_{m,D}^{(direct)} * s_m \right](t) + \left[h_{m,D}^{(indirect)} * s_m \right](t) \right\} + v_D(t) , \quad (3)$$

where * denotes the convolution, $h_{m,D}$ represents the filter between the source m and the microphone D (decomposed

into a direct and indirect paths), $v_D(t)$ is the intrinsic noise of the microphone D, and D stands for either a_n (or simply n), W, X, Y or Z.

In order to simplify (3), we make the assumption that the direct path is frequency-independent (which is not true in practice, because of directivity and radiation effects) so we can parameterize it by a simple delay $\tau_{m,D}$ and a gain $g_{m,D}$:

$$h_{m,D}^{(direct)} = g_{m,D} \cdot \delta(t - \tau_{m,D}) . \quad (4)$$

The parameters $\tau_{m,D}$ (resp. $g_{m,D}$) represent the delay (resp. the attenuation or amplification) of the m-th acoustic source captured by the main microphone for $D=W$, or by the n-th spot microphone for $D=n$. For $D=X,Y,Z$ the parameters $g_{m,D}$ represent the directional encoding of the m-th acoustic source:

$$\begin{cases} g_{m,X} = g_{m,W} \cdot \eta \cdot \cos\theta_m \cdot \cos\varphi_m \\ g_{m,Y} = g_{m,W} \cdot \eta \cdot \sin\theta_m \cdot \cos\varphi_m \\ g_{m,Z} = g_{m,W} \cdot \eta \cdot \sin\varphi_m \end{cases} . \quad (5)$$

Parameters Estimation

The estimation of the parameters (delay, position, gain) cannot rely on the unavailable source signals, but only on main and spot microphone signals. The first step is to align (time-synchronize) the different spot microphone signals with the main microphone signal. Then, the estimation of the position and the gain becomes possible. This parameter estimation is achieved frame by frame.

The delay estimation for each spot microphone will be achieved thanks to an analysis between it and the W omnidirectional component of the main microphone. On one hand, it is quite impossible to estimate $\tau_{m,W}$ and $\tau_{m,n}$ with the available signals. On the other hand, the necessary delay parameter is given by the following difference:

$$\tau_n = \tau_{m,W} - \tau_{m,n} , \quad (6)$$

that is the exact delay that we must apply to a spot microphone for synchronizing it with the main microphone.

We introduce the cross-correlation as a dot product between two time-shifted signals for a discrete and finite support [12]:

$${}_{\tau_1}\langle x | y \rangle_{\tau_2} = \sum_{k=K_1}^{K_2} x(k-\tau_1) y(k-\tau_2) , \quad (7)$$

where τ_1 and τ_2 are the values of time shifting in terms of samples, K_1 and K_2 the limits of a reference time segment (frame).

We also use the notation $\langle x | y \rangle_\tau = {}_0\langle x | y \rangle_\tau$, and by introducing the norm of a discrete signal: $\|x\|_\tau = \sqrt{\langle x | x \rangle_\tau}$, we note that $\|x\| = \|x\|_0$.

These definitions being introduced, we will show how to estimate the parameters: delay, position and gain.

We estimate the delay using a normalized cross-correlation function $C(\tau)$ between the signals $W(t)$ and $a_n(t)$. Since the spot microphone signal is more representative of the source to be mixed than the main microphone signal (which usually contains also concurrent sources), it is more relevant to compute correlation regarding a fixed portion of $a_n(t)$ while $W(t)$ is time-shifted by the opposite delay:

$$C(\tau) = \frac{\langle a_n | W \rangle_{-\tau}}{\|a_n\| \cdot \|W\|_{-\tau}} . \quad (8)$$

By replacing $W(t)$ using (3) and (4) we obtain:

$$C(\tau) = \frac{\langle a_n | g_{m,W} \cdot s_m(t - \tau_{m,W}) \rangle_{-\tau}}{\|a_n\| \cdot \|g_{m,W} \cdot s_m(t - \tau_{m,W})\|_{-\tau}} , \quad (9)$$

where $\tau_{m,W}$ is a delay between the source and the W omnidirectional component of the main microphone under the following assumptions:

- the indirect paths and intrinsic noises are neglected ($h_{m,W}^{(indirect)} = v_w(t) = 0$),

- only one source m is active at a given time interval ($M=1$).

It can be shown from (3) and (4) that (9) can be expressed as:

$$C(\tau) = \frac{\langle a_n | a_n(t + \tau_{m,n} - \tau_{m,W}) \rangle_{-\tau}}{\|a_n\| \cdot \|a_n(t + \tau_{m,n} - \tau_{m,W})\|_{-\tau}} . \quad (10)$$

Let $\tau_n = -(\tau_{m,n} - \tau_{m,W})$, (10) can be rewritten as:

$$C(\tau) = \frac{\langle a_n | a_n \rangle_{\tau_n - \tau}}{\|a_n\| \cdot \|a_n\|_{\tau_n - \tau}} . \quad (11)$$

It is obvious that this function reaches its maximum (and unit) value for $\tau = \tau_n$.

Based on this simplified assumptions we introduce the estimator $\tilde{\tau}_n$ associated to the parameter τ_n as the index corresponding to the maximum of the normalized cross-correlation function:

$$\tilde{\tau}_n = \underset{\tau}{\text{Argmax}} C(\tau) , \quad (12)$$

In practice, especially with periodic signals, the presence of other concurrent active sources ($M>1$) may disturb delay estimation (12) and raise secondary peaks of the cross-correlation function above the peak corresponding to the delay (**Figure 3** (top)). To provide better robustness we introduce a time smoothing of the normalized cross-correlation that enhances the function (8) for each i-th frame by using the FIR filter [12]:

$$C'_i(\tau) = \sum_{k=0}^{K} b_k C_{i-k}(\tau) . \quad (13)$$

where K is the order of filter, b_k denotes the filter's coefficients. In the case of simple mean function, $b_k=1/(K+1)$.

This method is efficient when the period of the signal changes. Indeed the peak corresponding to the delay remains stable and is preserved with time smoothing, while other peaks move (around the searched one) proportionally to the signal period and therefore are lowered by smoothing (**Figure 3** bottom).

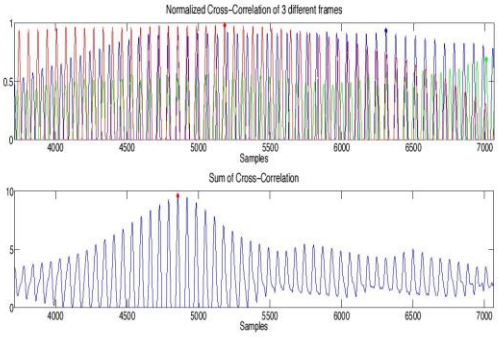

Figure 3: Example of one harmonic signal (top). Extended cross-correlation function (bottom).

Another smoothing that we use in practice to avoid memorizing several sets of cross-correlation values, is a first order autoregressive filtering of the cross-correlation function (11):

$$C_i'(\tau) = \alpha \cdot C_{i-1}'(\tau) + (1-\alpha) \cdot C_i(\tau) . \qquad (14)$$

This only requires memorizing the values of the smoothed cross-correlation of the previous frame $C'_{i-1}(\tau)$. α is a forgetting factor that can be made adaptive or fixed, which is the case for the experiment presented later.

In order to estimate the spatial information (azimuth $\tilde{\theta}_n$ and elevation $\tilde{\varphi}_n$), we use the estimated delay $\tilde{\tau}_n$ identified in (12) and the following dot products obtained from the three HOA components (X, Y, Z) in (2):

$$\begin{cases} \langle a_n | X \rangle_{-\tilde{\tau}_n} = \eta \cdot \langle a_n | W \rangle_{-\tilde{\tau}_n} \cdot \cos\theta_n \cdot \cos\varphi_n \\ \langle a_n | Y \rangle_{-\tilde{\tau}_n} = \eta \cdot \langle a_n | W \rangle_{-\tilde{\tau}_n} \cdot \sin\theta_n \cdot \cos\varphi_n \\ \langle a_n | Z \rangle_{-\tilde{\tau}_n} = \eta \cdot \langle a_n | W \rangle_{-\tilde{\tau}_n} \cdot \sin\varphi_n \end{cases} . \qquad (15)$$

Thus, the estimated azimuth $\tilde{\theta}_n$ is given by:

$$\tilde{\theta}_n = \operatorname{atan2}(\langle a_n | Y \rangle_{-\tilde{\tau}_n}, \langle a_n | X \rangle_{-\tilde{\tau}_n}) , \qquad (16)$$

and the estimated elevation $\tilde{\varphi}_n$ is given by:

$$\langle a_n | Z \rangle_{-\tilde{\tau}_n} = \eta \cdot \langle a_n | W \rangle_{-\tilde{\tau}_n} \cdot \sin\tilde{\varphi}_n \Rightarrow$$
$$\tilde{\varphi}_n = \arcsin\left(\frac{\langle a_n | Z \rangle_{-\tilde{\tau}_n}}{\eta \langle a_n | W \rangle_{-\tilde{\tau}_n}}\right). \qquad (17)$$

The estimated gain $\tilde{g}_n$ is simply obtained by the following ratio:

$$\tilde{g}_n = \frac{\langle a_n | W \rangle_{-\tilde{\tau}_n}}{\|a_n\|^2} . \qquad (18)$$

Performance Evaluation

For this experimental part, the length of the frame (the buffers of the dot products) is arbitrarily set to 512 samples at 48kHz, and the maximum delay is set to 7060 samples. The step between two successive frames is 256 samples (50% of the frame length). The forgetting factor α is set to 0.99, which corresponds to a convergence time (defined as the time required to cover 90% of the distance between two stationary situations after transition between them) of 1.22 second.

We evaluated the performance of the algorithm on two versions (without and with reverberation) of a sound scene composed of 8 acoustic sources (see **Table 1**) captured separately in an anechoic chamber (Mozart, An aria of Donna Elvira from the opera Don Giovanni, [14]). The first version is a mix of 8 acoustic sources without reverberation, using ideal HOA encoding. The second version of the scene makes use of Spatial Room Impulse Responses (SRIR) recorded with an Eigenmike in the "Opera de Rennes" for a variety of source positions, and converted in the HOA format. Each anechoic source signal is convolved with the SRIR associated to its position, before being summed into the global HOA mix. In this situation, simulated spot microphone signals remain anechoic and crosstalk-free.

The first difficulty is the presence of several acoustic sources playing notes simultaneously with harmonic relationships, leading to sharing common spectral components. The second problem is caused by the reverberation that introduces additional acoustic waves with exactly the same spectral content as the direct sound.

The angles and delays of **Table 1**, used for sound scene synthesis, derive from the analysis of the first wave front of the SRIR used for the reverberated scene.

Sources	Azimuth	Elevation	Delay	Gain	Color
Bassoon	-29	20	4934	1	
Clarinet	31	-5	4606	1	
Flute	27	-32	4914	1	
Contrabass	26	29	5085	1	
Voice	29	13	4746	0.3	
Violoncello	30	-24	4917	1	
Violin	-29	26	5077	1	
Viola	-22	-39	4898	1	

Table 1: Description of sound scene

For the sake of readability we draw the parameters estimations only for two acoustic sources (bassoon and voice) among the 8 sources that compose the scene. All estimations are related to frames containing significant energy (associated source is considered as active when its energy is above a threshold of -60 dB) to exclude irrelevant periods of signal and therefore avoid bad estimations and improve the robustness. We can observe (**Figure 4**) that all the parameters are very often well estimated. Variations in position estimation are often limited except in few places. However, the gain estimation is less robust than the position or delay. Like other estimators, the dot product in (18) is altered by the presence of concurrent sources in $W(t)$ that are partially correlated with $a_n(t)$. But unlike position estimators, there is no "counterweight" by another dot product that might be similarly impacted.

It is interesting to note that with a forgetting factor α set to 0.9, leading to errors in delay estimation (**Figure 5**), the position estimation is not disturbed. Indeed the delay estimation error typically corresponds to an integer number of signal periods, and the portion of $W(t)$ that "wrongly" matches with the reference block of $a_n(t)$ looks very like the "right" one, thus the dot products are only slightly affected.

The analysis of the second scene, with reverberation (**Figure 6**), shows large errors in position estimation that can be explained by the secondary acoustic waves (the reflections), which disturb the signal. Moreover, the delay is also estimated with errors but the estimation converges towards a value close to the target one.

Discussion

The analysis module should be proposed to end-users (sound engineers) as a tool with an interface assisting them during the mix. One simple possible interface could show the variation of possible values around the supposed target value (as in **Figure 4** top and mid), and help users to take a final decision about the right parameter and realize the sound tuning.

The procedure of assistance module usage for the sound engineer could be described in few steps:

- Predefine the initial state of a sound scene: information about sources (e.g. maximal distance between spot and main microphones), environment estimation (e.g. distance between spot or main microphones and a reverberant obstacle);
- Observe the results of parameters estimation in one of the graphical or statistical representation (e.g. evolution of parameter along time or histogram of the values);
- Adjust the control (tuning) based on results of each estimated parameter.

Both of the evaluated scenes are using fixed sources and associated spot microphones. However, it should be noted a variety of other cases which are not covered yet in this paper.

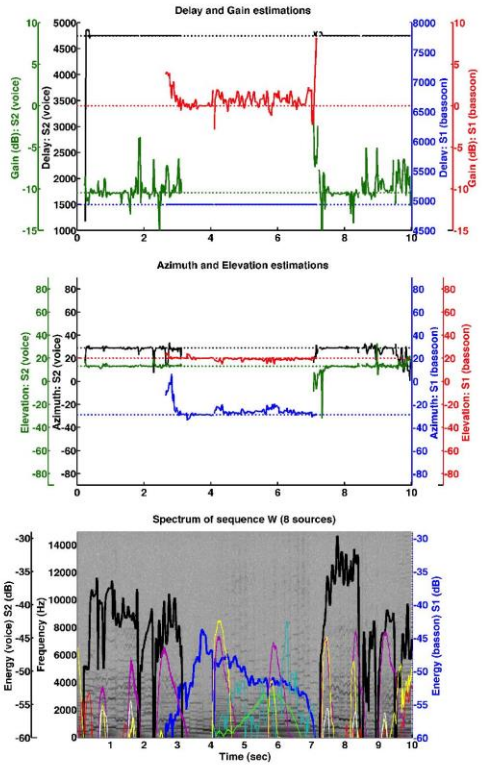

Figure 4: Estimation of delay, gain (top) and position (mid) of bassoon and voice (target values plotted with dotted lines). Bottom: spectrogram of signal W of the main microphone, with level indication of the 8 acoustic sources that compose the scene (without reverberation). See **Table 1** for colors associated to each source.

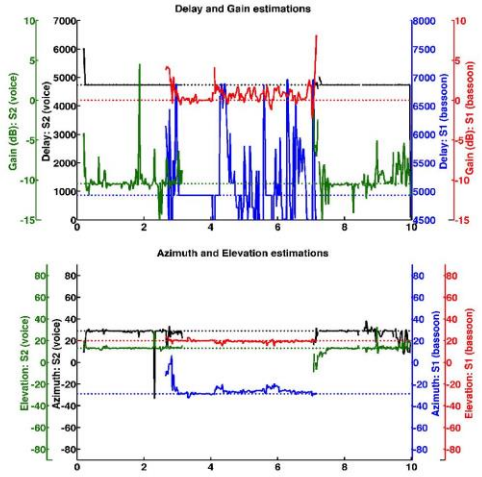

Figure 5: Similar as **Figure 4** with an error in delay estimation for bassoon.

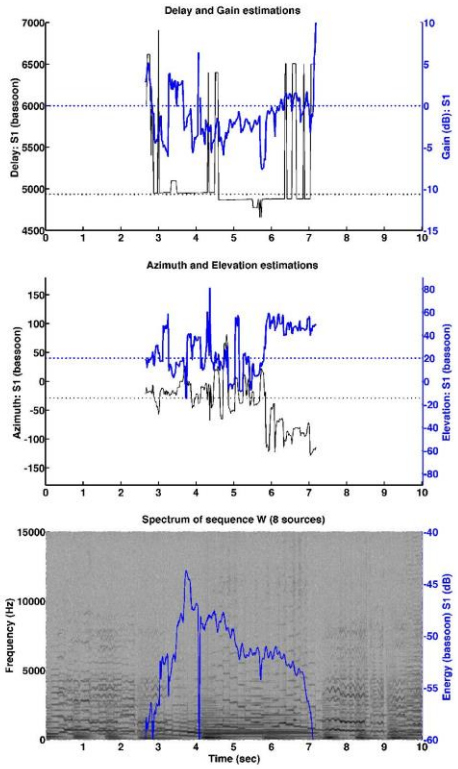

Figure 6: Delay, gain and position estimations for the bassoon. Similar as **Figure 4** (scene composed of 8 sources) but with reverberation.

One of them is a sound scene with moving sources (for example, the play of an actor with Lavalier microphone). In this case we can denote the advantage of frame-by-frame analysis that provides numerical values corresponding to estimated parameters at each frame. Therefore, the values of each frame represent the evolution of source parameters in time (distance (i.e. delay) and gain between main and spot microphones; elevation and azimuth of the sources).

Another case is a sound scene with moving sources in reverberant environment. The secondary acoustic waves (the reflections), which change the propagation trajectory with the source position, could produce large errors in parameters estimation.

Thus, additional information about a sound scene could be useful for analysis, e.g. a measure of reverberant environment (room or concert hall for example) as well as approximated zone of source position in space. Thanks to this information, the analysis module can reduce the possible data set and, at the same time, improve the algorithm robustness.

Conclusion

In this paper we presented a typical workflow for mixing hybrid 3D content ("natural 3D" content captured by a main HOA microphone, and "artificial 3D" content spatially encoded in HOA from sources simultaneously captured by monophonic spot microphones). In this context, we proposed an algorithm providing estimated parameters (delay, position, gain) that could be used to process the spot microphone signals during the mix. This algorithm makes use of spatial encoding equations in the HOA format, and relies on some assumptions on the available signals. We evaluated it on controlled artificial sound scenes with different conditions. The first results seem encouraging since the graphs show good performances in presence of several acoustic sources. However, the algorithm is less robust with complex cases like in presence of reverberation, which effect has to be further evaluated by controlling its amount and characteristics. Ongoing work focuses on robustness improvement by searching methods for an adaptive forgetting factor and by associating confidence indicators to the estimated parameters (allowing to update the parameters or to keep the previous ones). We will also address complexity reduction. As next steps, we plan to analyze real, recorded scenes including moving sources. Analysis of real recorded scenes such as orchestra will also raise the "crosstalk" issue that has not been handled in this paper. A particular and probably difficult case concerns the capture of an instrumental section (e.g. a group of violins playing the same score) by a unique spot microphone. We also plan to support stereo spot microphones.

We plan to implement the algorithm as a VST plugin [15] for Digital Audio Workstations to allow its usage by sound engineers. We also plan to support stereo spot microphones.

References

[1] ISO/IEC 23008-3/DIS, 3D Audio MPEG standard. URL: http://mpeg.chiariglione.org/standards/mpeg-h/3d-audio/ dis-mpeg-h-3d-audio

[2] M. A. Gerzon: Periphony: With-Height Sound Reproduction, J. Audio Eng. Soc., vol. 21, no. 1, pp. 2–10, Feb. 1973

[3] J. Daniel: Représentation de champs acoustiques, application à la transmission et à la restitution de scènes sonores complexes dans un contexte multimedia. PhD thesis, Paris 6, 2000

[4] SoundField Microphones and Processors. URL: http://www.tslproducts.com/soundfield-type/soundfield-microphones/

[5] Eigenmike® HOA microphone. Digital signal processing, acoustics and product design. URL: http://www.mhacoustics. com/

[6] Brahma Ambisonic Microphone (http://www.embrace-cinemagear.com/brahma-ambisonic-microphone.html)

[7] Core Sound TetraMic® Sound Microphone. URL: http://www.core-sound.com/TetraMic/1.php

[8] C.A. Englebert: Réalisation d'une réverbération à convolution et mise en place de méthodes de mixage pour la technique de spatialisation sonore 'High Order Ambisonics (HOA), Research paper, Salle Pleyel, 2013

[9] Sébastien Moreau: Étude et réalisation d'outils avancés d'encodage spatial pour la technique de spatialisation sonore Higher Order Ambisonics: microphone 3D et contrôle de distance, PhD Thesis, 2006

[10] Blue Ripple Sound Limited, specialist in Spatial Audio technology. URL: http://www.blueripplesound.com/

[11] Matthias Kronlachner, ambiX v0.2.3 – Ambisonic plug-in suite. URL: http://www.matthiaskronlachner.com

[12] Sophocles J. Orfanidis: Introduction to signal processing. Rutger University. URL: http:// www.ece.rutgers.edu/~orfanidi/intro2sp/orfanidis-i2sp.pdf

[13] Conservatoire National Supérieur de Musique et de Danse de Paris - Formation Supérieure aux Métiers du Son. URL: http://www.fsms.fr/

[14] Aalto University, School of Science, Department of Media Technology. Anechoic recordings of symphonic music. URL: https://mediatech.aalto.fi/en/ research/virtual-acoustics/research/acoustic-measurement-and-analysis/85-anechoic-recordings

[15] Steinberg, Third-party developer support. URL: https://www.steinberg.net/en/company/developers.html