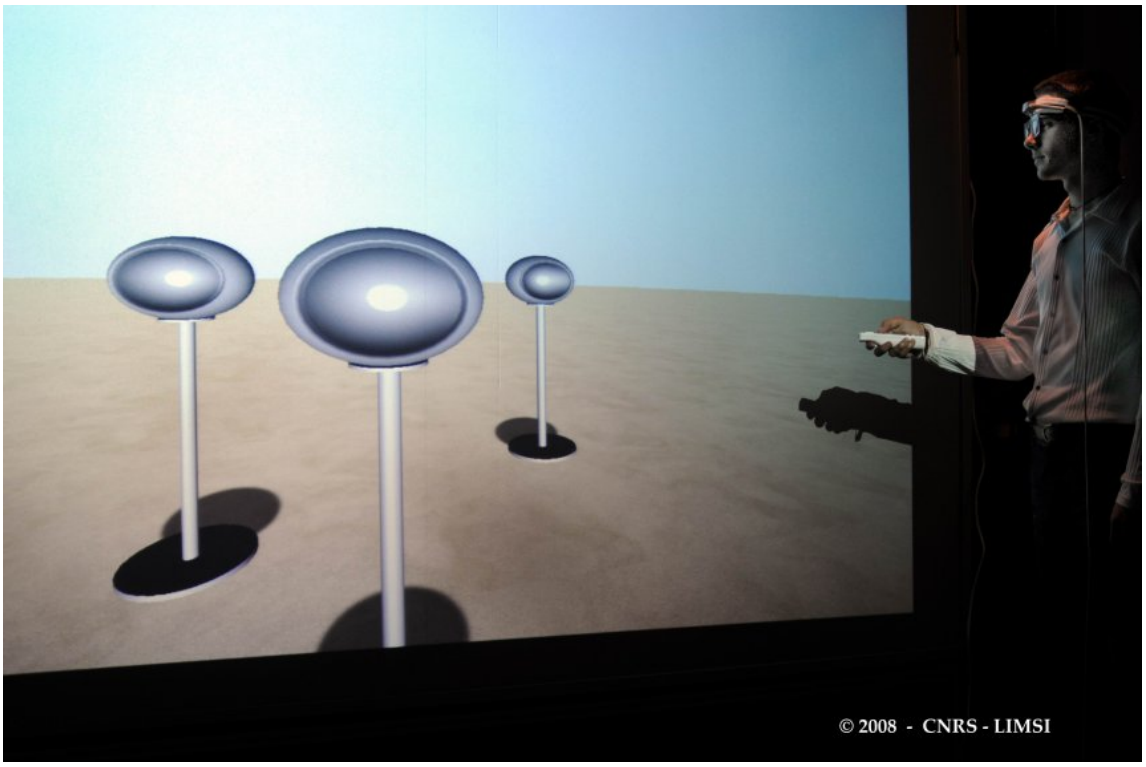
Master de Sciences et Technologies de l'UPMC
spécialité Mécanique et Ingénierie des Systèmes
Parcours Acoustique Musicale
Acoustique, Traitement du Signal et Informatique Appliqués à la Musique

# Cohérence spatiale des informations visuelles et auditives dans un environnement de réalité virtuelle




Marc Rébillat

*Mars-Août 2008*

Brian F.G. KATZ, LIMSI-CNRS
Etienne CORTEEL, *sonic emotion*

# Table des matières

## Résumé

Dans la majorité des interfaces de réalité virtuelle conçues à ce jour, l'intégration des modalités visuelle et auditive n'est pas optimisée. D'une part, l'accent est souvent mis sur la modalité visuelle, au détriment de la restitution audio. D'autre part, les interfaces visuelles nécéssitent presque toujours la présence d'un écran de projection. Il y'a donc une forte concurrence spatiale entre les positions optimales des emetteurs acoustiques et de l'écran de projection, qui nécéssite des compromis.

Ainsi, nous proposons ici un nouveau concept de design d'interfaces audio-visuelles baptisé SMART-$I^2$ (Spatial, Multi-user, Audio-visual, Real-Time, Interactive Interface). Au sein de ce dispositif, le rendu visuel est assuré via la technologie de *stéréoscopie avec tracking*, et le rendu audio avec la technologie de *wave field synthesis* (WFS). L'obstacle posé par les difficultés de positionnement de l'écran et des émetteurs acoustiques est levé par l'utilisation de la technologie des *multi-actuators panel's* (MAP's) de grandes dimensions qui permettent le design à la fois d'un banc de haut-parleurs et d'un écran, sans contraintes spatiales particulières.

Pour la réalisation de MAP's de grandes dimensions, un encastrement latéral est nécéssaire pour le maintien du panneau en position. Une étude acoustique a donc été menée pour essayé de quantifier l'influence que pouvait avoir l'encastrement d'un MAP sur la qualité de son rayonnement (réponse fréquentielle, directivité et non-linéarités). Ainsi, une modélisation, couplée à une méthode de mesure ont été développées et appliquées dans ce contexte. Puis des indices globaux, sensés par rapport aux objectifs, ont été définis pour permettre la comparaison entre les différents types d'encastrement à l'étude. L'encastrement retenu à la suite de cette étude a ensuite été celui appliqué lors de la construction du SMART-$I^2$.

Enfin, une validation du rendu audio-visuel prodigué par le SMART-$I^2$ a été conduite. Le choix a été fait de coupler une étude objective uniquement acoustique et une étude subjective du rendu audio-visuel. L'étude objective, utilisant une tête artificielle et des mesures *in situ* des réponses impulsionnelles des émetteurs, a permis de voir quelles erreurs étaient commises sur les indices acoustiques nécéssaires à la localisation dans la zone de restitution. L'étude subjective, a permis de quantifier avec quelle précision des sujets étaient capables de localiser une source audio parmi un panel de sources visuelles. Les deux études ont montré que le rendu audio-visuel proposé par le SMART-$I^2$ était précis et convaincant en terme de cohérence spatiale.

## Remerciements

Je tiens a remercier tout d'abord mes encadrants Etienne Corteel et Brian Katz pour m'avoir fait confiance, encadré, et donné la chance de réaliser ce prototype. Je remercie aussi Yves Maire sans qui le SMART-$I^2$ n'aurait jamais tenu debout. Merci aussi, aux membres de l'équipe Interfaces Hommes-Machines (Tifanie Bouchara, Matthieu Courgeon, Rami Ajaj, Christian Jacquemin, ...) pour leur aide qui me fut précieuse pour m'y retrouver dans les méandres de l'informatique graphique.

# Chapitre 1

# Introduction
# Le projet SMART-$I^2$ dans son contexte

Dans ce chapitre, le projet SMART-$I^2$ (acronyme pour "Spatial, Multi-user, Audio-visual, Real Time, Interactive Interface") est présenté ainsi que le contexte qui lui a permis de voir le jour.

## 1.1 Le LIMSI et *sonic emotion*

Ce stage est le fruit d'une collaboration scientifique menée par deux entités : le LIMSI (Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur, rattaché au CNRS) et *sonic emotion*, entreprise suisse spécialiste du son 3D.

Le LIMSI est un laboratoire au sein duquel un large spectre disciplinaire est couvert, allant du "thermodynamique au cognitif". Ce stage a été hébergé par le groupe "Perception située" du LIMSI. Ce groupe a pour objectif l'étude de la perception dans ses diverses modalités pour le développement de systèmes de perception artificielle et d'interfaces expressives. Il traite des modalités visuelle, auditive et gestuelle, considérées dans les 4 dimensions de l'espace et du temps. En particulier, l'utilisation du son spatialisé dans de nouvelles applications de réalité virtuelle est l'un de ses centres d'intêret.

L'entreprise Suisse *sonic emotion* est à l'heure actuelle un acteur majeur du marché mondial du son 3D. Elle propose des solutions technologiques pour la spatialisation du son pour des applications variées (projets artistiques, installations académiques, loisirs grand public, ...). Pour l'instant restée concentrée sur la spatialisation du son, elle est aussi intéressée par ses applications possibles dans un contexte audio-visuel.

## 1.2 Présentation du projet

Traditionnellement, les environnements audio-visuels de réalité virtuelle sont conçus autour de deux structures bien distinctes : les écrans de projection pour la restitution visuelle et les hauts parleurs pour la restitution audio. Les développements de ces dernières années ont eu tendance à favoriser séparément le rendu 3D de l'une ou l'autre de ces modalités. De plus, ce découplage engendre des incohérences de restitution car l'écran n'est jamais acoustiquement transparent, et des gênes pour les utilisateurs qui nécéssairement voient les haut-parleurs dans le cadre de situations d'immersion. Ainsi, les dispositifs existants ne restituent à l'utilisateur qu'une sensation de présence limitée.

La notion de présence est ici interprêtée comme "l'illusion perceptive de non-médiation". Alors, suivant cette interprêtation, un dispositif de rendu audio-visuel se doit d'apparaitre comme invisible aux sens de l'utilisateur. Il doit être perçu uniquement comme une fenêtre grande ouverte, au travers de laquelle les utilisateurs et le contenu de la scène partagent le même monde physique. Par conséquent, une attention particulière doit être apportée au réalisme perceptuel, qui désigne l'habilité du dispositif à produire des scènes sensoriellement convaincantes. C'est autour de cet axe de réflexion que le SMART-$I^2$ a été conçu.

Dans les dispositifs ou un rendu audio global est proposé, préférentiellement à l'utilisation d'écouteurs (comme les environnements de type CAVE par exemple), il y a toujours concurrence entre les positions optimales de l'écran et des vidéo-projecteurs, et celles des haut-parleurs. Alors, des compromis d'organisation doivent être trouvés pour réaliser effectivement le rendu audio-visuel. Ainsi, les haut-parleurs sont souvent placés derrière ou sur les bords de l'écran de projection. Ces choix posent des problèmes pour les technologies de restitution sonore spatialisées qui nécéssitent un fin contrôle du champ de pression dans la zone de restitution. Le réalisme perceptuel n'est donc pas optimisé dans ces dispositifs car ce contrôle y reste limité.

Une autre façon de restituer spatialement le son est d'utiliser un rendu binaural via des écouteurs. La qualité de ce type de rendus est étroitement liée à l'utilisations de filtrages individualisés pour chaque utilisateur. De plus l'utilisation d'écouteurs réduit le degré d'immersion et l'interactivité possible entre les utilisateurs. En effet, idéalement les utilisateurs ne devraient pas être contraints par le système de rendu sonore pour communiquer. C'est pourquoi ce type de rendu n'a pas été retenu dans le cadre du projet SMART-$I^2$.

Nous nous sommes donc tournés vers un nouveau type de haut-parleurs pour répondre à nos attentes : les Multi-Actuator Panel's (MAP's). Il s'agit de panneaux rigides au dos desquels sont fixés des excitateurs et qui rayonnent de façon analogue à des bancs de hauts parleurs. Pour l'instant cette technologie a été utilisée pour construire des haut-parleurs de dimensions relativement réduites (133 cm $\times$ 67 cm). Dans le cadre du projet SMART-$I^2$, le but est de réaliser d'un seul tenant à la fois un banc de haut-parleurs et un écran de projection en utilisant cette technologie. Les dimensions seront donc nécéssairement plus conséquentes et un système de maintien du panneau devra être développé de façon à optimiser le rayonnement de la structure. Ainsi l'influence des conditions d'encastrement sur le rayonnement des MAP's est un point auquel il faudra prêter attention.

L'utilisation de ce dispositif à la fois comme un banc de haut-parleurs et comme écran de projection nous affranchis ainsi a priori des problèmes crées par la dissociation audio/vidéo évoquée précédement. À ce MAP sont ensuite associées de façon cohérentes des technologies de restitutions visuelle et sonores qui permettent de plonger l'utilisateur dans un monde virtuel. Pour la restitution sonore, la "Wave Field Synthesis", qui permet de recréer un champ sonore dans le plan horizontal dans une large zone de restitution a été choisie. Le rendu visuel est assuré par la stéréoscopie trackée, qui envoie à chacun des yeux de l'utilisateur les images de la scène, relativement à la position qu'il occupe.

L'association de ces deux technologies est réalisée dans le but de restituer non seulement les indices audio-visuels statiques de localisation, mais aussi des indices dynamiques de localisation, comme l'effet de parralaxe. Le mouvement des utilisateurs dans la zone de restitution leur permet en effet d'estimer la position des objets par les mouvements relatifs qu'ils effectuent les uns par rapport aux autres. Et aussi bien la Wave Field Synthesis que la stéréoscopie trackée restituent naturellement cet effet. Ainsi, le choix de ces deux technologies s'avère prometteur pour la réalisation du SMART-$i^2$.

## 1.3 Enjeux du stage

Ce stage consiste donc en la conception, l'optimisation et la validation d'un tel dispositif. Les deux axes d'étude qui ont été abordés, relativement à la problématique présenté précédement, sont les suivants :

1. Pour répondre aux exigences de la projection, et pour des questions pratiques, les panneaux servant d'écran et de haut-parleur ont besoin d'être peint et maintenus latéralement. Il est donc important d'étudier les conséquences acoustiques de ces choix de façon à les faire en optimisant le rendu sonore.
2. La finalité de ce dispositif est de stimuler la perception humaine de façon convaincante. Il est donc nécéssaire d'évaluer perceptivement et objectivement la qualité du rendu offert à l'utilisateur. Dans ce stage, seule la qualité de localisation des objets audio-visuels sera évaluée et les autres dimensions (timbre, ...) seront laissées de côté.

## 1.4 Organisation du mémoire

Le mémoire qui suit est organisé autour de quatre chapitres de façon à mettre en avant la logique du stage par rapport aux enjeux énoncés précédement. Le premier, qui s'achève ici est une brève introduction du contexte qui a permis au SMART-$I^2$ d'être réalisé. Le second est une étude bibliographique nécéssaire à la compréhension des technologies utilisées et à la compréhension des motivations du projet. Puis, dans un troisième temps, une étude acoustique focalisée sur l'optimisation du rayonnement de Multi-Actuators Panel's est présentée. Ensuite, autour d'un quatrième chapitre, la réalisation à proprement parler du prototype ainsi que la procédure de validation sont expliqués. Ce chapitre fait l'objet d'une publication de conférence et sera présenté lors de la $125^{eme}$ convention de l'Audio Engineering Society (AES) à San Francisco. Enfin deux annexes cloturent ce mémoire. L'une propose des figures supplémentaires pour une meilleure compréhension de l'étude acoustique menée au chapitre 3 et l'autre détaille la façon dont la cohérence entre le monde physique et le monde virtuel a été assurée.

# Chapitre 2

# Etude bibliographique

Ce chapitre propose une étude bibliographique succinte des capacités humaines dans les domaines de l'ouie et de la vision, et des technologies de reproduction associées (i.e. la Wave Field Synthesis et la Stéréoscopie Trackée). Enfin, la notion de transparence inhérente à tout environnement de réalité virtuelle est explicitée.

## 2.1 Audition Spatiale

Dans un environnement donné, plusieurs indices permettent au cerveau de localiser une source sonore dans l'espace. Le repère naturel dans lequel nous percevons le son est le repère de coordonnées sphériques dont notre tête est l'origine (cf. figure 2.1). Nous allons donc étudier l'aptitude humaine à localiser une source selon les trois coordonnées de ce repère égocentrique (distance, azimuth et élévation ou cite). Pour de plus amples informations concernant la perception spatiale du son, on pourra se référer à [Bla97].

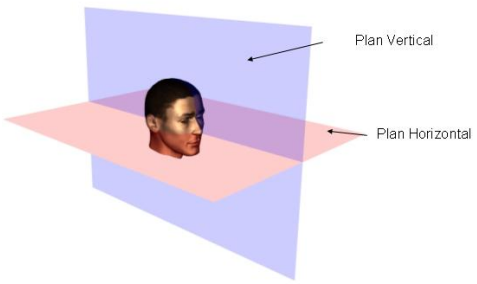



FIG. 2.1 – Repère sphérique associé à la tête de l'auditeur.

### 2.1.1 Localisation dans le Plan Horizontal

Le plan horizontal est le plan le plus important de l'espace perceptif sonore. La localisation dans ce plan est obtenue par la comparaison des champs sonores arrivant à chacune des oreilles. Trois indices, repartis en deux catégories ITD (Interaural Time Differences) et ILD (Interaural Level Differences) nous servent à déterminer la direction de provenance des sources dans le plan horizontal :

1. **Les différences interaurales de temps (ITD)**
   (a) *La différence de phase* entre les signaux sonores arrivant sur chacune des oreilles. L'oreille y est sensible surtout pour les fréquences inférieures à 1500 Hz.
   (b) *La différence de temps d'arrivée de l'enveloppe* du signal sonore au niveau de chaque oreille. On considère que cet indice est surtout utilisé à partir de 1500 Hz.

2. **Les différences interaurales de niveau sonore (ILD)** dûes à l'atténuation de l'onde sonore avec la distance et à la présence de la tête. Cet indice est valable sur toute la gamme de fréquences audibles, mais la diffraction des ondes sonores par la tête ne se manifeste qu'à partir de 1500 Hz environ.

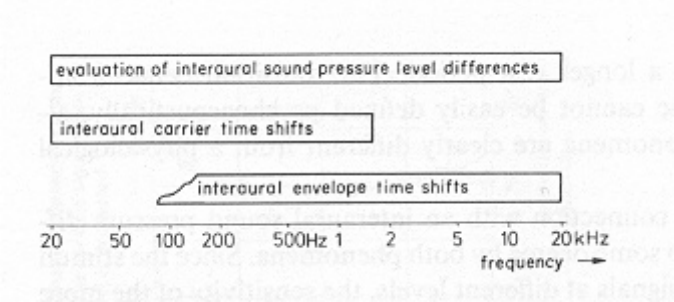



FIG. 2.2 – Importance relative des différents indices en fonction de la fréquence (d'après [Bla97]).

Le système d'audition humain emploie pour la localisation de sources dans le plan horizontal au moins deux de ces indices. Il semble que la différence inter-aurale de temps soit l'indice le plus influent pour la localisation de sons ayant une part importante de leur énergie en dessous de 1500 Hz [Wig92]. À partir de ces trois indices, nous sommes donc en mesure de définir de façon assez précise (i.e. avec une précision de l'ordre de 3° [Bla97]) la provenance d'une source sonore dans le plan horizontal.

### 2.1.2 Localisation dans le Plan Vertical

Dans le plan vertical les champs sonores incidents sur chacune des oreilles étant égaux, c'est alors la forme du pavillon qui nous permet de situer les sources. En effet le pavillon agit comme un réflecteur de son, un résonateur se comportant différemment en fonction de la direction d'incidence des ondes sonores. D'autre part, le torse et les épaules de l'auditeur, par les réflexions qu'ils apportent aident aussi à la localisation dans le plan vertical.

Ainsi, un filtrage fréquentiel est réalisé par nos pavillons qui dépend de la direction de provenance de la source sonore. Les fonctions de transferts entre une direction donnée et les tympans sont mesurables et appelées "Head Related Transfert Function" (HRTF). Le contenu spectral du son initial se retrouve donc modifié s'il vient d'au dessus de nous ou d'ailleurs dans le plan vertical. C'est alors l'apprentissage inconscient et individuel des filtrages fréquentiels réalisés par notre pavillon en fonction de la direction d'incidence du son qui nous permet d'identifier la direction de provenance d'une source sonore dans le plan vertical.

Il semble quand même que notre capacité de localisation dans le plan vertical soit beaucoup moins précise que dans le plan horizontal, et surtout qu'elle dépende assez fortement de notre connaissance du contenu spectral de la source. La précision de localisation varie de 4° pour un bruit blanc à 17° pour la voix d'une personne que l'on ne connait pas (cf. [Lok07]).

### 2.1.3 Perception de la Distance

Pour ressentir l'éloignement ou la proximité d'une source sonore, nous faisons appel à trois indices :

1. **L'intensité** : Nous sommes conscients que le niveau sonore d'une source décroit au fur et à mesure que l'on s'en éloigne. Ainsi, l'intensité est un indice très fin pour juger de variations de distance par rapport à une source ou la distance séparant deux sources, mais non pour estimer la distance absolue à laquelle se situe une source. De plus lorsque deux sources sonores sont très proches de nous (à moins d'un mètre) les différences de niveau sonore arrivant aux deux oreilles sont perceptibles et donc l'ILD est un indice très pertinent pour estimer la distance entre ces sources.

2. **La part de champ réverbéré dans le son perçu** : Dans une salle, au fur et à mesure que l'on s'éloigne d'une source, nous percevons les réflexions multiples des ondes issues de cette source plutôt que les ondes directement issues de la source. Le rapport entre l'énergie du champ direct et l'énergie du champ réverbéré est donc un bon indicateur de distance absolue à laquelle se situe une source sonore.
3. **Le spectre de la source** : Pour des sources qui nous sont connues, leur contenu spectral nous renseigne sur leur proximité. Par exemple, à niveau sonore égal, la voix de quelqu'un qui chuchote nous paraitra toujours plus proche que celle de quelqu'un qui parle normalement.

L'oreille humaine est donc un organe assez performant en terme de localisation d'évènements sonores, surtout dans le plan horizontal.

## 2.2 La Wave Field Synthesis

La Wave Field Synthesis est la technologie de rendu spatialisé du son qui a été utilisée pour le SMART-$I^2$. Ses fondements théoriques ainsi que ses limitations pratiques sont ici présentés.

### 2.2.1 La WFS en théorie

La Wave Field Synthesis s'appuie sur le principe de Huygens-Fresnel pour restituer un champ de pression donné dans une zone étendue de l'espace. Ce principe peut être formulé comme il suit :

*Chaque point de l'espace est considéré comme indépendent. Si un point M reçoit une onde d'amplitude $a(M,t)$, alors on peut considérer qu'il réémet lui-même une onde sphérique de même fréquence, même amplitude et même phase que l'onde reçue. Les fronts d'onde créés par les sources secondaires interfèrent ensuite entre eux. Au lieu de considérer que l'onde progresse de manière continue, on décompose sa progression en imaginant qu'elle progresse de proche en proche (voir figure 2.3).*

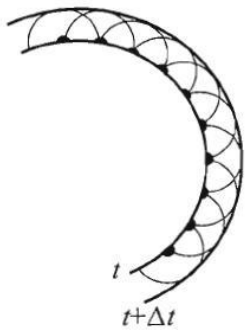



FIG. 2.3 – Illustration du principe de Huygens Fresnel (extrait de [Bru04]).

Ainsi, considérons un volume de reproduction $\Omega_R$, et un volume $\Omega_\Psi$ ou sont situées les sources, tout deux séparés par une surface fermée $\partial\Omega$. il est alors possible de reproduire dans le volume $\Omega_R$ n'importe quel champ sonore issu de sources contenues dans $\Omega_\Psi$, pour peu que l'on connaisse à tout instant l'amplitude et la phase du champ incident en tout point de la surface $\partial\Omega$ séparant les deux volumes (voir figure 2.4 pour les notations).

On peut alors déduire de ce principe une formulation intégrale, dans le domaine fréquentiel, du champ de pression règnant dans l'espace de restitution $\Omega_R$ qui est donnée par l'équation 2.1 ( voir [Ber93] ).

$$P_R(\overrightarrow{r_R},\omega) = \iint_{\partial\Omega} \left( \quad P_\Psi(\overrightarrow{r_S},\omega)\overrightarrow{\nabla}[G(\overrightarrow{r_R},\overrightarrow{r_S})] - G(\overrightarrow{r_R},\overrightarrow{r_S})\overrightarrow{\nabla}[P_\Psi(\overrightarrow{r_S},\omega)] \quad \right).\overrightarrow{dS} \tag{2.1}$$

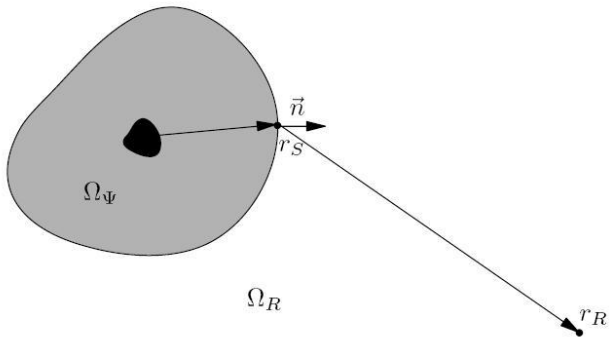



FIG. 2.4 – Définition de l'espace des sources $\Omega_\Psi$ et de l'espace de reproduction $\Omega_R$ (extrait de [Cor04]).

avec :
- $P_R(\overrightarrow{r_R}, \omega)$ : la transformée de Fourier de la pression au point de réception.
- $P_\Psi(\overrightarrow{r_S}, \omega)$ : la transformée de Fourier de la pression émise par la source et reçue en un point de la surface $\partial\Omega$.
- $G(\overrightarrow{r_R}, \overrightarrow{r_S}) = \frac{\exp -jk||\overrightarrow{r_R} - \overrightarrow{r_S}||}{4\pi||\overrightarrow{r_R} - \overrightarrow{r_S}||}$ qui est la fonction de Green, en 3 dimensions, dans le domaine fréquentiel.

Et cette intégrale vaut par définition :

$$\begin{aligned} P_R(\overrightarrow{r_R}, \omega) &= P_\Psi(\overrightarrow{r_R}, \omega) \quad \text{pour tout les points de } \Omega_R \\ &= 0 \quad \text{pour tout les points de } \Omega_\Psi \end{aligned}$$

L'équation 2.1 montre donc que le champ rayonné dans l'espace de restitution peut être considéré comme la superposition :
- d'un champ rayonné par une distribution surfacique de monopôles alimentés par la composante normale du gradient du champ de pression de la source.
- d'un champ rayonné par une distribution surfacique de dipôles alimentés par l'amplitude du champ de pression de la source.

De plus, dans le cas particulier ou la surface $\partial\Omega$ est un plan infini, les contributions des monopôles et dipôles sont en phase dans $\Omega_R$ et en opposition de phase dans $\Omega_\Psi$. L'information contenue par chacune des distributions surfaciques est donc redondante. Il est alors possible de réduire l'intégrale à l'un ou l'autre des types de sources. Les éléments électro-acoustiques actuels étant plutôt des monopôles, c'est la représentation monopolaire que l'on choisira.

Alors on obtient l'équation 2.2, aussi appelée "Equation de Rayleigh 1", qui décrit le champ de pression dans l'espace de restitution comme étant issu d'une répartition surfacique continue de sources monopolaires disposées sur un plan infini.

$$P_R(\overrightarrow{r_R}, \omega) = -2 \iint_{\partial\Omega} G(\overrightarrow{r_R}, \overrightarrow{r_S}) \overrightarrow{\nabla} [P_\Psi(\overrightarrow{r_S}, \omega)] \quad .\overrightarrow{dS} \tag{2.2}$$

A partir de cette formulation du principe de Huygens-Fresnel il est donc théoriquement possible de reproduire dans une zone étendue de l'espace un champ sonore donné à partir d'une répartition surfacique continue de monopôles.

### 2.2.2 La WFS en pratique

En pratique, il n'existe pas de répartition surfacique plane, infinie et continue de monopôles contrôlable en tout point. Nous sommes donc amenés à réaliser des approximations pour mettre en oeuvre le principe décrit précédement (cf. [Cor04]).

#### Réduction à une ligne de sources secondaires

La réduction de la distribution surfacique de sources à une distribution linéique n'est pas obligatoire, mais est motivée par les arguments suivants :

- La tête de l'auditeur est dans la majorité des cas dans le plan des sources qu'il écoute.
- La localisation dans le plan horizontal est la plus précise perceptivement, donc la plus importante à restituer.
- On réduit de beaucoup la complexité du système en réduisant le nombre de sources.

Pour réaliser cette réduction, on considère qu'un point M de la ligne finale devra émettre une contribution égale à la somme des contributions des sources qui lui sont perpendiculaires. Ainsi, en utilisant l'approximation de la phase stationnaire, on peut calculer la pression que doit rayonner chaque point de la ligne de sources secondaires [Cor04].

Ceci à deux conséquences :

- Dorénavant, le champ de pression rayonné est à symétrie cylindrique, autour de l'axe que constitue la ligne sur laquelle sont réparties les sources.
- Le passage de la répartition surfacique à une répartition linéique fait intervenir un coefficient d'amplitude dépendant de la distance à laquelle on se trouve du banc de sources. L'atténuation du champ sonore réel n'est donc bien restituée que sur une ligne parallèle à la ligne que constituent les sources (profondeur d'écoute).

#### Passage à un segment

Pour des raisons d'encombrement, il n'est pas non plus possible d'avoir une ligne infinie de sources. Par conséquent, la zone où se situe les sources est réduite à un segment, ce qui pose deux problèmes :

- **Diffraction** : Le segment de sources sonores peut être vu comme une fenêtre acoustique ouverte entre deux espaces sonores. Etant donné la taille des longueurs d'onde associées aux phénomènes acoustiques (de quelques centimètres à plusieurs mètres !), il y a donc apparition d'un phénomène de diffraction par cette fenêtre qui parasitera le champ acoustique que l'on cherche à obtenir.
- **Limitation de zone d'écoute** : La limitation de la distribution de sources à un segment va limiter la taille de la zone d'écoute dans laquelle il est possible de restituer correctement les sources (voir figure 2.5).

#### Discrétisation des sources secondaires

La dernière approximation est que l'on ne sait pas commander de segment continu de sources sonores. Nous sommes donc contraints de réaliser un échantillonnage spatial des sources, comme montré sur la figure 2.6.

Une fréquence de repliement spatial $f_{al}$ existe donc et est donnée par l'équation 2.3.

$$f_{al}(\Psi, r) = \frac{1}{\max_{L\in[1..N]} \Delta t_L(\Psi, r)} \tag{2.3}$$

ou :

- $\Psi$ désigne la source à laquelle on se réfère.
- $r$ désigne la position d'écoute par rapport au centre du dispositif.
- $L$ désigne une source parmis les $N$ présentes.

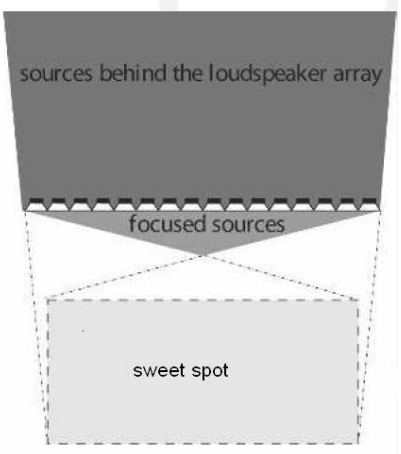



FIG. 2.5 – Positions possibles des sources en fonction de la taille du banc de haut parleurs et de la zone d'écoute (extrait de [Cor06ME]).

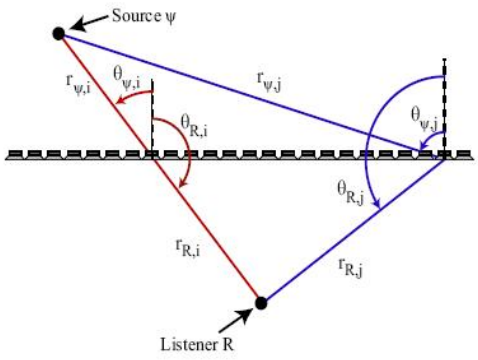



FIG. 2.6 – Echantillonnage spatial des sources secondaires, d'après [Cor04].

– $\Delta t_L(\Psi, r) = t_{L+1}(\Psi, r) - t_L(\Psi, r)$ représente la différence de temps d'arrivée entre deux sources adjacentes.

Pratiquement, un écart de 15 à 20 cm entre deux sources consécutives est suffisant pour avoir une fréquence de repliement acceptable ($f_{al} \approx 1500$ Hz) et ainsi garantir une bonne localisation des sources sonores synthétisées. En effet, avec cette fréquence d'aliasing, les différences de phase (ITD) pour des fréquences inférieures à 1500 Hz sont bien restituées ce qui donne perceptivement une localisation plutôt précise. Néanmoins, l'influence perceptive de cette fréquence d'aliasing n'est pas encore bien comprise, d'autant plus que cette fréquence varie en fonction de la position de l'auditeur.

Sur la figure 2.7, la réponse impulsionnelle temporelle du banc de haut parleur fait apparaître d'abord un premier front d'onde ou tous les haut parleurs contribuent en phase et qui correspond à celui synthétisé, et ensuite apparaissent les contributions individuelles de chacun des haut parleurs. Au niveau de la réponse impulsionnelle spatiale, on retrouve la même chose en voyant des fronts d'ondes secondaires se propageant derrière le front d'onde principal.

### 2.2.3 Multi Actuator Panels (MAP's)

Pour réaliser les sources sonores et ainsi mettre en oeuvre la W.F.S., on peut utiliser soit des hauts parleurs électrodynamiques classiques en réseau soit une autre technologie de reproduction sonore appelée "Multi Actuator Panels" (cf. [Boo04]) dérivée des haut-parleurs à modes distribués (DML : Distributed Modes Loudspeakers).

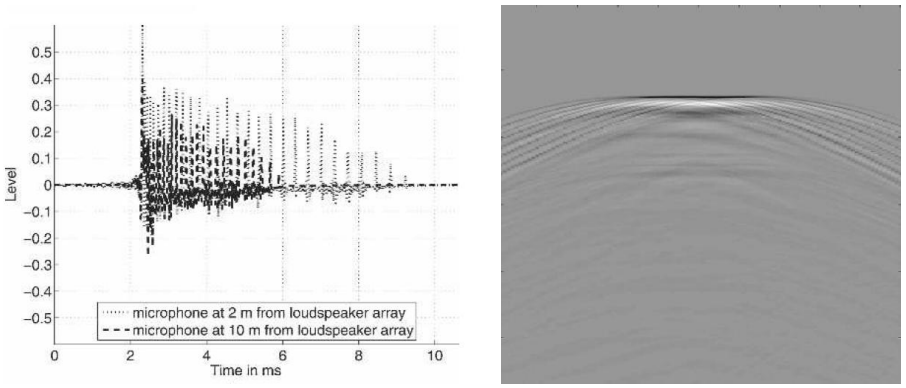



FIG. 2.7 – Illustration des effets de l'échantillonage spatial. A gauche : Réponse impulsionnelle temporelle d'un banc de sources sonores (extrait de [Cor06ME]). A droite : Réponse impulsionnelle spatiale d'un banc de sources sonores reproduisant une onde plane (extrait de [Boo04]).

Il s'agit de panneaux d'un matériau léger et rigide au dos desquels viennent se fixer des excitateurs électrodynamiques (cf. figure 2.8). Alors, si le matériau choisi présente un suffisament bon coefficient d'amortissement interne, il est possible de considerer que les rayonnements de chacun des excitateurs sont indépendants, et donc ne s'influencent pas l'un l'autre. Ainsi on dispose de l'équivalent d'un banc discret de haut-parleurs, mais qui ne sont pas visuellement gênants.

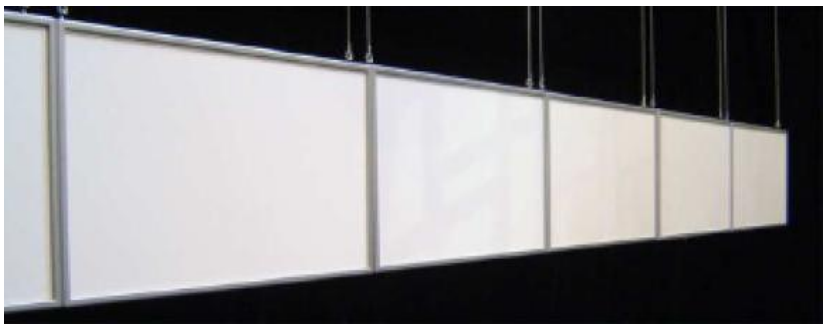


FIG. 2.8 – Alignement de MAP's.

Dans [Cor07], une étude a été menée pour comparer les restitutions offertes par les haut-parleurs et les MAP's dans le cadre de la Wave Field Synthesis :

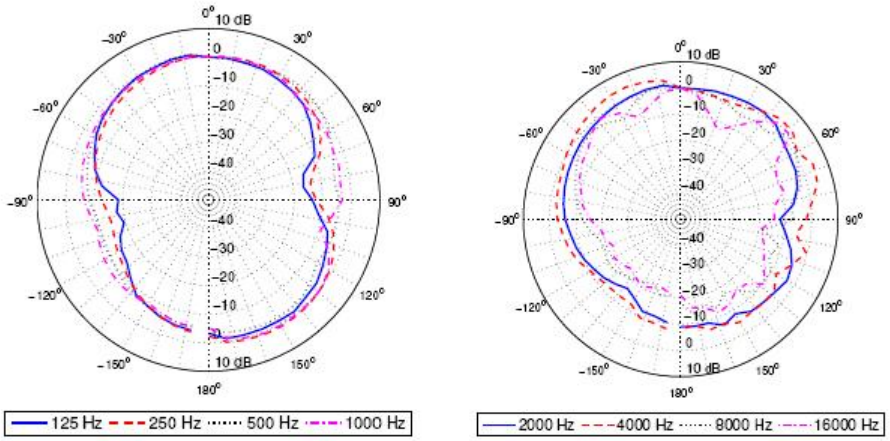
1. Du point de vue de la directivité, les MAP's présentent un comportement plus compliqué que les haut-parleurs, ce qui est a priori compensable par des techniques d'égalisation. D'autre part, à l'opposé des haut-parleurs qui deviennent directifs en hautes fréquences, les MAP's restent assez omnidirectionnels en hautes fréquences.
2. Une autre particularité des MAP's est de présenter un comportement diffus, du à une densité modale élevée en hautes fréquences. Ce comportement diffus s'avère gênant pour la cohérence spatiale du rayonnement en basses fréquences, mais présente aussi l'avantage de limiter les incohérences liées à la fréquence d'aliasing spatial.

Cette étude émet en conclusion l'hypothèse que le comportement diffus des MAP's aurait pour conséquence d'augmenter la transparence de la reproduction sonore.

### 2.2.4 Egalisation par inversion multicanal pour la Wave Field Synthesis

Le dernier point à prendre en compte pour la restitution de l'espace sonore est la façon dont rayonnent les émetteurs dans l'espace en fonction de la fréquence. En effet la théorie développée

précédement est basée sur l'hypothèse que le champ sonore est reproduit à l'aide d'une distribution de sources monopolaires parfaitement omnidirectionnelles et à réponse fréquentielle plate. Or, les haut-parleurs ou les MAP's ne présentent jamais de caractéristiques aussi idéales, comme on peut le constater pour les MAP's sur la figure 2.9.




Directivité en basses fréquences d'un haut-parleur Map

Directivité en hautes fréquences d'un haut-parleur Map

FIG. 2.9 – Directivité des Maps par bandes d'octave en normalisant la réponse fréquentielle par rapport à la réponse à 0°. Mesures réalisées sur des panneaux de 40 cm par 60 cm, (extrait de [Cor07]).

Alors, pour essayer de minimiser l'influence de ces défauts, des techniques d'égalisations ont été employées. La technique d'égalisation classique consiste à considérer séparément chacun des émetteurs et à essayer de compenser sa réponse fréquentielle moyennée dans l'espace. Avec cette technique, on ne prend que très peu en compte les caractéristiques de rayonnement des émetteurs. Ce n'est donc pas un moyen très efficace de contrôler le champ sonore dans l'espace.

Une autre technique d'égalisation pour la W.F.S. utilise le filtrage inverse multicanal (voir [Cor06ME]). Cette technique considère le problème comme un système à plusieurs entrées (les actuateurs) et plusieurs sorties (des microphones disposés à des points ou l'on souhaite contrôler le champ de pression). Connaissant les entrées, la mesure des sorties donne accès à une matrice de transfert entrée/sortie. D'autre part, connaissant la réponse idéale désirée en chacun des points de sortie il est alors possible de calculer l'erreur entre le modèle et les mesures et par les moindres carrés et estimer ainsi les filtres nécéssaires pour minimiser cette erreur (cf. figure 2.10).

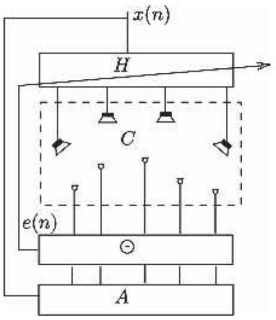



FIG. 2.10 – Schéma de principe de la méthode d'égalisation multicanal (extrait de [Cor06ME]).

Alors, le champ de pression est contrôlé en chacun des points de sortie, mais pas du tout en dehors

de ces points. Donc pour étendre la validité spatiale de cette méthode, une représentation physique du champ à contrôler (type condition sur une surface limite pour rester cohérent avec la WFS) est choisie. Maintenant c'est le champ au niveau de chaque point de cette surface qu'il est nécéssaire de contrôler, ce qui assure a priori une validité spatiale complète du contrôle en vertu de l'équation 2.1.

Cependant, pour réaliser ceci, il faudrait pouvoir disposer d'un plan de micro continu et infini. Il faut donc envisager les mêmes approximations au niveau du système de mesure que pour le système de rendu :

- Le plan de microphones est réduit à une ligne.
- La ligne de microphones est dicrétisée.
- La ligne de microphones est réduite à un segment.

Ainsi, c'est seulement la composante normale du champ de pression au niveau d'un segment discret que l'on contrôle réellement, ce qui assure quand même en pratique une validité spatiale étendue de la zone de contrôle, mais en dessous d'une fréquence d'aliasing relative au segment de microphones. Cette méthode d'égalisation réussit alors à réduire de façon efficace les artefacts de rendu audio, en particulier dans les cas ou les actuateurs sont des MAP's et donc ont des diagrammes de directivité assez complexes.

## 2.3 Restitution Tridimensionnelle de la Vision

Dans le cadre du projet, nous souhaitons reproduire un environnement de réalité virtuelle qui soit audio-visuel. LA W.F.S. propose un rendu 3D du son et nous allons voir maintenant une technologie proposant un rendu visuel 3D convaincant.

### 2.3.1 Oeil et acuité visuelle

L'oeil est un organe très complexe, et nous n'allons ici préciser que quelques caractéristiques fondamentales de la perception du monde que nous offrent nos deux yeux. On pourra se référer à [Fuc01] pour plus d'informations concernant cette partie.

La fonction des yeux est de canaliser la lumière, de longueur d'onde comprise entre 400 et 700 nanomètres, émise ou réfléchie par un objet pour créer une image nette qui s'imprime sur la partie de l'oeil recouverte de récepteurs sensoriels : la rétine. L'oeil est constitué d'une succession de milieux transparents qui jouent le rôle d'une lentille convergente dont la focale globale peut varier par modification de la courbure du cristallin (voir figure 2.11).

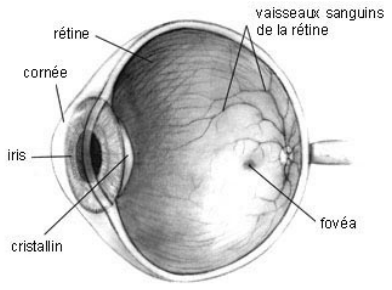



FIG. 2.11 – Vue en coupe de l'oeil humain (extrait de [Fuc01]).

La rétine est constituée de deux sortes de photo-récepteurs :

- **Les cônes** : concentrés sur la fovéa, ils sont sensibles à la longueur d'onde.
- **Les batônnets** : présents presque partout sur la rétine, ils sont au contraire très peu sensibles à la couleur.

Même si les batonnets sont beaucoup plus nombreux que les cônes ($120.10^6$ contre $6.10^6$) ce sont les cônes qui contribuent en majorité à l'information transmise au cerveau. Par conséquent l'acuité visuelle n'est pas homogène sur tout le champ de vision.

Pour un oeil normal :
- L'acuité monoculaire est un cône de 2° d'ouverture.
- La valeur minimale de l'angle sous lequel deux points distincts sont perçus dépend du type de stimulus mais est de l'ordre de 1'.
- Le champ de vision horizontal est de 90° côté tempe et 50° côté nez. Ce qui créé une zone de recouvrement d'environ 100° dans laquelle la vision est binoculaire.
- Le champ de vision horizontal est de 45° vers le haut et de 70° vers le bas.

L'oeil est donc capable d'une très forte précision, mais dans une zone très limitée de l'espace. Pour compenser ceci, il à la possibilité de bouger d'environ 15° dans les orbites à une vitesse maximale de 600°/s. Par conséquent en réalité virtuelle, un bon rendu visuel couvre un champ de vision horizontal de 170° et un champ de vision vertical de 145°. La persistence rétinienne étant d'environ 23 Hz, il est nécéssaire que les images soient projetés à une fréquence supérieure ou égales pour ne pas être perçues séparément.

### 2.3.2 Perception de la profondeur

Contrairement à l'intuition, la perception de la profondeur ne découle pas uniquement de la vision binoculaire. En effet, de part le fait qu'en vision naturelle les yeux et la tête bougent, de nombreux indices monoculaires servent aussi à la perception de la profondeur et donc de la troisième dimension spatiale.

#### Indices monoculaires

Les indices monoculaires pour la perception de la profondeur sont assez nombreux, les plus importants sont :
- **Les dimensions relatives des objets** : Chaque objet envoie au fond de l'oeil une image dont les dimensions sont proportionelles à celle de l'objet et décroissantes en fonction de la distance par rapport à l'oeil. Le cerveau connaissant les dimensions "normale" des objets réels peut donc en déduire leur distance approximativement.
- **Le parallaxe du au mouvement** : Quand un observateur se déplace, les images rétiniennes d'objets immobiles ont des mouvements relatifs dépendant des distances de ceux-ci. Ceci nous permet d'apprécier la sensation de profondeur d'une scène.

#### Indices binoculaires

Ces indices sont issus de la convergence des axes optiques de chacun des yeux et du recouvrement partiel de leurs champs visuels. En effet cela créé deux vues légèrement différentes d'une même scène permettant d'en extraire une information de profondeur.

Si les axes optiques de nos yeux convergent vers le point F (voir figure 2.12) alors le point A renvoie sur chacun des yeux des images différentes : $A_g$ et $A_d$. La disparité, dans l'hypothèse ou les axes (GD) et (AF) sont orthogonaux, est l'angle :

$$d = \beta - \alpha = 2.\widehat{DF_dA_d}$$

L'analyse de cette disparité par le cerveau nous renseigne sur la profondeur relative de l'objet A par rapport à F. Cependant, étant donné que cet angle ne prend des valeurs notables que si les objets sont près des yeux alors cet indice est efficace surtout à des distance inférieures à quelques mètres.

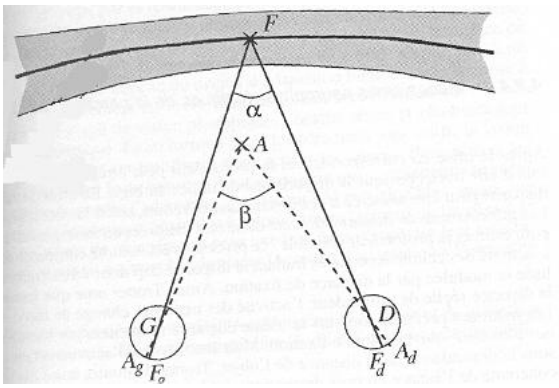



FIG. 2.12 – Perception de la profondeur par la disparité (extrait de [Fuc01]).

La disparité la plus petite perceptible est de $65.10^{-3}$ radians. Cela nous donne la capacité de percevoir les objets dans des plans éloignés de $\Delta_r = 10^{-3}r^2$, si $r$ est la distance à laquelle nos axes optiques convergent.

### 2.3.3 La stéréoscopie

Pour avoir un rendu visuel convaincant et cohérent, il faut d'une part être en mesure de générer une scène visuelle ou tous les indices monoculaires soient correctement rendus, mais aussi d'amener à chaque oeil l'image qui lui est destinée (ce qu'on appelle la stéréoscopie). Pour ce faire, plusieurs techiques existent. La stéréoscopie active, basée sur le principe du multiplexage temporel, utilise des lunettes à volets, synchronisées avec le vidéo-projecteur. Le volet de chaque oeil n'est ouvert que lorsque l'image projetée lui est destinée. La stéréoscopie passive utilise les propriétés de polarisation de la lumière pour envoyer à chaque oeil l'image qui lui est destinée. Ce type de projection consiste à projeter sur un écran métallisé (et qui donc conserve la polarisation) simultanément les deux images destinées à l'utilisateur mais polarisées à 90° l'une par rapport à l'autre. L'utilisateur doit ensuite regarder cet écran avec des lunettes présentant les mêmes type de filtres polarisants pour que les yeux perçoivent uniquement l'image qui leur est destiné (voir figure 2.13). Cette technique à l'avantage de moins fatiguer l'utilisateur que la stéréoscopie active mais par contre, les filtres polarisants n'étant pas parfaits, des images fantômes peuvent apparaitre.

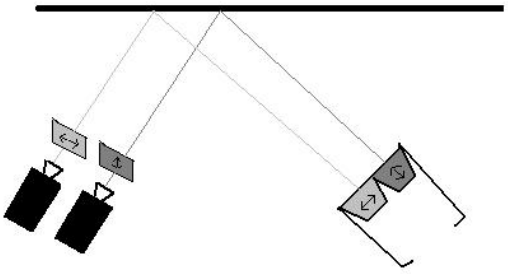


FIG. 2.13 – Principe de la projection polarisée

### 2.3.4 La nécéssité du tracking

Un rendu visuel statique induit des mouvements non désirés des objets lorsque l'utilisateur évolue dans la zone de restitution (cf. figure 2.14). C'est ce que l'on appelle la pseudoscopie et cela est du au fait que la projection soit plane, et que le point de vue proposé à l'utilisateur soit fixe .

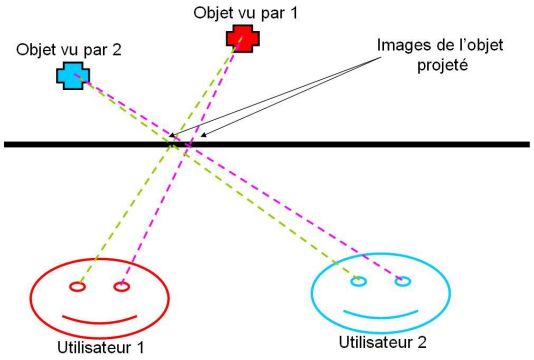



FIG. 2.14 – Mouvements pseudoscopiques : les utilisateurs 1 et 2 perçoivent l'objet visuel à des positions différentes.

Pour réduire cet inconvénient, il est nécéssaire de suivre en temps réel la tête de l'utilisateur pour projeter à chaque instant le bon point de vue, d'ou l'utilisation d'un trackeur.

## 2.4 Transparence de l'interface audio-visuelle

Il est évident que quelles que soient les techniques de reproduction qui sont utilisées, leur fidélité par rapport à la realité sera toujours limitée. Un système idéla se doit d'être transparent pour que l'utilisateur se consacre entièrement à son immersion dans le monde virtuel et ne soit pas perturbé par des artefacts liés aux moyens de restitution. Ainsi, les limitations technologiques risquent d'introduire des artefacts perceptifs qui nuiront à l'immersion totale de l'utilisateur dans le monde qui lui est proposé. La transparence de restitution est donc une notion qui a pour but d'estimer l'écart perceptif entre le but qui est recherché et le but qui est atteint (voir figure 2.15). Cette notion peut s'inclure plus généralement dans la notion de présence [Lom97], et peut s'apparenter à la fois au réalisme percéptuel et au degré d'immersion.

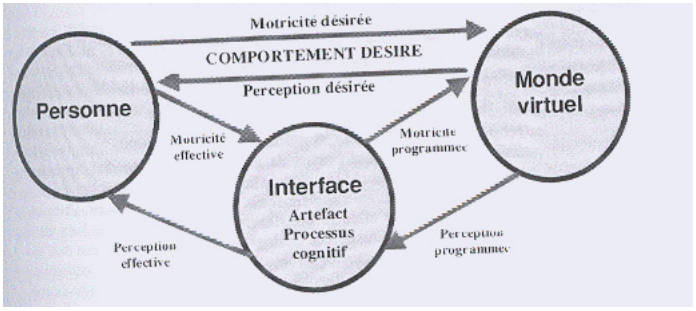



FIG. 2.15 – Illustration de la notion de transparence en réalité virtuelle (d'après [Fuc01]).

Il est donc important de voir quels critères influent sur la transparence de l'interface et d'essayer de maximiser l'immersion des utilisateurs dans le système en miniminsant les artefacts perceptifs.

# Chapitre 3

# Optimisation acoustique d'un Multi-Actuator Panel

Préliminairement à la réalisation du protoype du SMART-$I^2$, une étude centrée sur l'optimisation du rayonnement acoustique d'un Multi-Actuator Panel (MAP) a été effectuée. Ce chapitre présente la démarche et les choix réalisés pendant cette étude ainsi que les résultats obtenus.

## 3.1 Caractérisation acoustique du système

### 3.1.1 Présentation des systèmes à l'étude

Les MAP's que nous cherchons à caractériser sont réalisés avec un petit panneau de matériau acoustique de 40 cm × 60 cm × 4 mm auquel est attaché un excitateur, comme présenté sur la figure 3.1.

FIG. 3.1 – Vue de dessous du MAP caractérisé lors de l'étude.

Le constructeur ne nous a fourni que très peu de données concernant ce matériau. Nous connaissons uniquement sa masse surfacique qui est de 520 $g/m^2$. Nous avons donc préféré réaliser une caractérisation expérimentale de ce système plutôt qu'une modélisation accompagnée de simulations.

Nous avons choisi d'étudier le rayonnement acoustique de 4 systèmes, dont 3 utilisent ce panneau :

1. *Haut-parleur de référence* : Il s'agit d'un haut-parleur électrodynamique classique, qui nous sert de système de référence, car son comportement est plutôt bien connu et qu'il est supposé présenter un rayonnement de bonne qualité (voir figure 3.3).
2. *MAP libre* : C'est un MAP constitué du panneau de (40 cm × 60 cm × 4 mm) auquel est attaché un excitateur, comme présenté sur la figure 3.1. Ce MAP est en régime de fonctionnement que l'on qualifiera de "libre" car ses bords sont laissés libres, sans aucune contraintes de maintien.

3. *MAP avec encastrement partiel* : Ce MAP est constitué du même panneau que le système précédent mais avec un encastrement partiel, sur 50% du pourtour du panneau. Cet encastrement est réalisé à l'aide d'une mousse d'isolation phonique réalisant l'interface entre le panneau et des profilés d'encastrement en aluminium (cf. figure 3.2)
4. *MAP avec encastrement total* : Ce MAP est l'analogue du système précédent sauf que cette fois-ci l'encastrement est réalisé sur tout le pourtour du panneau.

FIG. 3.2 – Les profilés d'encastrement en aluminium utilisés pour maintenir le MAP, à gauche, le MAP une fois encastré, au centre, et le détail des profilé et de la mousse à droite.

Ces quatre systèmes nous permettent d'avoir un système de référence ainsi qu'un panel de conditions d'encastrements de type dissipatif. Alors il est possible de réaliser une comparaison entre les systèmes dans le contexte définie précédement.

### 3.1.2 Objectifs

Les MAP's sont des structures qui ont donc déja été étudiées auparavent dans un contexte de reproduction sonore [Boo04, Kus06, Cor07, Pue08]. Cependant, les non-linéarités presentes dans le rayonnement de telles structures et l'influence de l'encastrement sur leur rayonnement restent des points très peu abordés par la littérature.

Étant donné que les MAP's sont avant tout des haut-parleurs, il est pourtant important de caractériser les non-linéarités qu'ils rayonnent et de voir quels sont les paramètres disponibles pour les minimiser. D'autre part, pour la réalisation du SMART-$I^2$, les MAP's utilisés vont être de grande taille (266 cm $\times$ 200 cm) et donc il va être nécéssaire de les maintenir verticalement. L'influence de la structure d'encastrement est donc aussi une question importante à étudier et les résultats obtenus sur le petit MAP vont conditionner le design de la structure de maintien du grand MAP.

Alors, les objectifs de cette étude sont :
- De quantifier l'ordre de grandeur et la localisation des non-linéarités rayonnées par un MAP.
- D'étudier l'influence acoustique de l'encastrement du panneau sur la réponse fréquentielle et la directivité .
- De réaliser les choix adaptés pour l'encastrement du MAP de grande dimensions.

### 3.1.3 Setup Expérimental

Pour réaliser cette étude, nous disposons d'une pièce dont toutes les parois (hormis le sol) sont recouvertes de mousse absorbante, ce qui permet ainsi de limiter les réflexions, et est efficace au moins au dessus de 100 Hz.

La chaine de mesure utilisée est constituée de 4 micros DPA 4006, d'un pré-amplificateur pour les microphones, d'un amplificateur audio pour le MAP, et d'une carte d'acquisition externe (cf. 3.3). Les gains des 4 microphones ont été calibrés avant la campagne de mesures de façon à ce que les niveaux de sortie de chacun d'eux soient identiques.

FIG. 3.3 – Présentation du setup expérimental disponible au LIMSI. A gauche, l'amplificateur audio, le pré-amplificateur pour les microphones et la carte d'acquisition audio. A droite, l'antenne acoustique et le système de repérage spatial.

Ces 4 microphones sont maintenus côte à côte et espacés de 10 cm les uns des autres, formant ainsi une antenne acoustique. Le système à tester est alors placé au centre de la pièce, la face emettrice dirigée vers le plafond et surélevé par rapport au sol de façon à se rapprocher le plus possible du rayonnement en champ libre. Un système de repérage spatial permet ensuite, en déplaçant l'antenne de microphones, de quadriller sur 84 points un plan de 1.2 m × 1.05 m parallèle au sol et situé à 30 cm de la face émettrice. L'échantillonnage spatial de ce plan est constitué de 7 lignes espacées de 15 cm comportant chacune 12 points de mesure répartis tous les 10 cm (cf. figure 3.4).

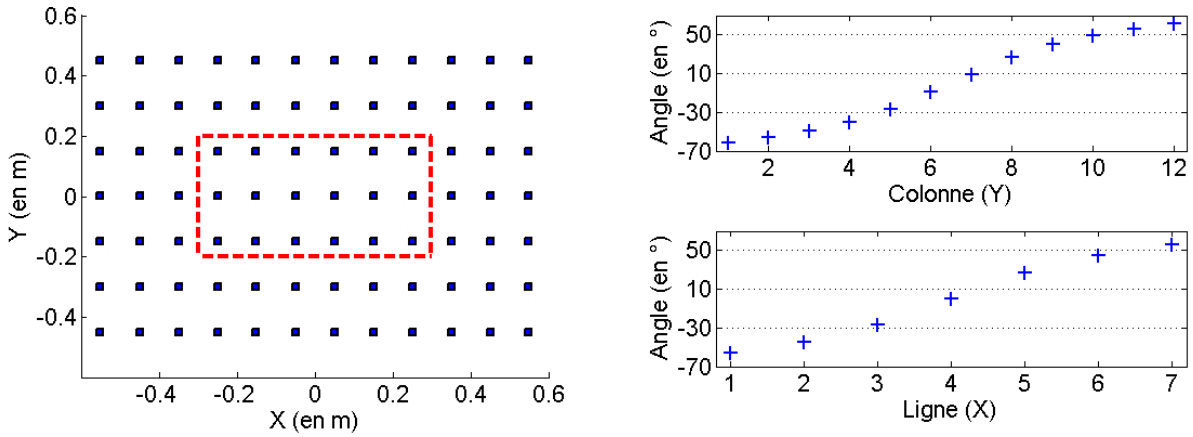



FIG. 3.4 – A gauche, positions des points de mesure par rapport au MAP. A droite, résolution angulaire offerte selon X et selon Y par les 84 points de mesure, depuis le centre du MAP.

Ainsi, avec ce dispositif, il est possible de caractériser le rayonnement d'une source sonore dans deux directions orthogonales de l'espace et avec une résolution angulaire de l'ordre de 10° (de -60° à 60° dans les deux directions).

## 3.2 Représentation non-linéaire du système

L'analyse non-linéaire d'un système suppose de mettre en oeuvre un outil permettant de caractériser les non-linéarités. Dans cette partie la méthode que nous avons utilisé pour cette étude est décrite.

### 3.2.1 Modélisation

Pour caractériser le système, il a été nécéssaire de choisir un modèle de non-linéarités. Parmi les différents modèles existants, le modèle de Hammerstein en cascade a été retenu [Ham30, Wes95]. Ce modèle propose de considérer le système non linéaire à étudier comme l'association de systèmes non-linéaire sans mémoire (produits instantanés indépendant de la fréquence) suivis de filtrages linéaires. Avec ce modèle, le système est par ailleurs supposé invariant dans le temps. Une représentation schématique en est donnée sur la figure 3.5.

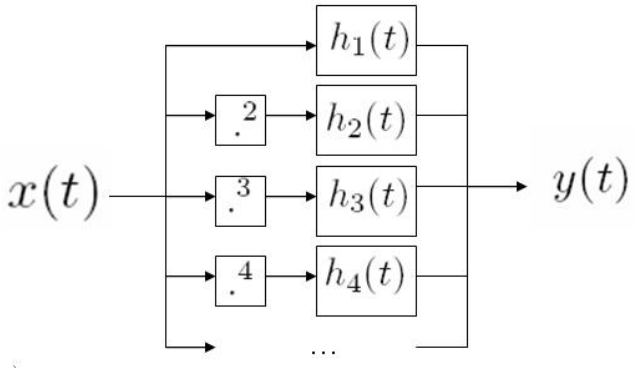



FIG. 3.5 – Vue schématique par blocs d'un système d'entrée $x(t)$ et de sortie $y(t)$ représenté avec le modèle d'Hammerstein en cascade. Les blocs "$.^n$" symbolisent les produits instantanés non-linéaires et sans mémoire.

Nous appelerons les fonctions $h_n(t)$ réponses impulsionnelles d'ordre n et leur transformées de Fourier, $H_n(f)$ les réponses fréquentielles d'ordre n. Mathématiquement, il est alors possible d'exprimer dans le domaine temporel la sortie $y(t)$ en fonction de l'entrée $x(t)$ comme une somme de produits de convolution (cf. équation 3.1). Cette relation peut aussi s'écrire dans le domaine fréquentiel d'une façon équivalente (cf. équation 3.2), avec $X_n(f) = TF[x^n(t)](f)$.

Cette modélisation met donc à disposition une représentation d'un système non-linéaire assez facile à manier, car elle ne repose que sur une famille de fonctions ne dépendant que d'une seule variable : les réponses fréquentielles (ou impulsionnelles, de façon équivalente) d'ordre n : $\{H_n(f)\}_{n\in\mathbb{N}*}$ (ou $\{h_n(t)\}_{n\in\mathbb{N}*}$).

$$y(t) = \sum_{n=1}^{+\infty} \int_{-\infty}^{+\infty} x^n(\tau) h_n(t-\tau) d\tau \tag{3.1}$$

$$Y(f) = \sum_{n=1}^{+\infty} H_n(f) X_n(f) \tag{3.2}$$

Ce modèle est une version simplifiée du modèle de Volterra valable pour tous les systèmes faiblement non-linéaire (i.e. dont la relation entrée/sortie est développable en séries entières) [Rug81, Has99]. En effet, dans la théorie de Wiener-Volterra, le noyau de Volterra d'ordre n (l'analogue de la fontion de transfert d'ordre n) est une fonction de n variables $H_n(f_1, ..., f_n)$, et décrit toutes les interactions non-linéaires possibles entre les différentes composantes de fréquences présentes dans le signal d'entrée. Dans le cas présent, le modèle d'Hammerstein ne permet pas une représentation aussi complète de ces interactions car la réponse fréquentielle d'ordre n n'est fonction que d'une seule variable. Les hypothèses supplémentaires introduites par cette représentation au niveau de la modélisation physique ne sont cependant pas encore bien cernées dans la littérature. Néanmoins ce modèle a déja été utilisé pour pré-compenser les non-linéarités crées par un haut-parleur électrodynamique [Las05], mais aussi pour décrire des phénomènes non-linéaires tel que les réflexes musculaires [Cho06] ou encore dans des applications de codage de la voix [Tur03].

### 3.2.2 Lien avec le taux de distorsion harmonique

La façon la plus classique et la plus intuitive de représenter les non-linéarités introduites par un système est d'utiliser le taux de distorsion harmonique (TDH). Considèrons un signal d'entrée harmonique $x(t) = E\exp(j2\ pift), E \in \mathbb{R}$ et la sortie lui correspondant, dans le cas d'un système non-linéaire $y(t) = \sum_{n=1}^{+\infty} S_n \exp(j2\pi nft + \phi_n), \forall n : (S_n, \phi_n) \in \mathbb{R}^2$. Alors une définition du taux de distorsion harmonique du système, à la fréquence $f$ et au niveau d'excitation $E$, est donné par l'équation 3.3.

$$TDH(f,E) = \sqrt{\sum_{n=2}^{+\infty} (\frac{S_n}{S_1})^2} \tag{3.3}$$

L'évaluation complète de la distorsion introduite par un système nécéssite donc une mesure pour chaque couple (fréquence, niveau d'entrée). De plus, une représentation du TDH dans le plan (f,E) est obtenu de façon discrète seulement.

Considérons alors l'équation 3.4 qui formalise la sortie $Y(f)$ d'un système représenté par un modèle de Hammerstein et soumis à la même entrée que précédement : $x(t) = E\exp(j2\ pift), E \in \mathbb{R}$.

$$y(t) = \sum_{n=1}^{+\infty} E^{(n-1)} |H_n(nf)| \exp[j2\pi nft + \arg(H_n(f))] \tag{3.4}$$

Alors le taux de distorsion harmonique, à la fréquence $f$ et au niveau d'excitation $E$ et donné par l'équation 3.5.

$$TDH(f,E) = \sqrt{\sum_{n=2}^{+\infty} \frac{|H_n(nf)|^2}{|H_1(f)|^2} E^{(2n-2)}} \tag{3.5}$$

Cette écriture du TDH permet alors plusieurs commentaires :

1. Avec ce modèle, le TDH dépend explicitement du niveau d'entrée E. Cela est cohérent avec le fait que les réponses fréquentielles d'ordre n sont indépendantes du niveau ou elles ont été mesurées dans la théorie de Wiener-Volterra (cf. [Has99]).
2. La connaissance du niveau de mesure et de chacune des réponses fréquentielles sur la bande fréquentielle $[f_A, f_B]$ permet une estimation continue dans le plan $(f, E)$ du TDH entre les fréquences $f_A$ et $f_B$ pour tout $E$.

Ainsi, cette modélisation permet une interprétation physique des phénomènes non-linéaires observés assez intuitive, de par sa proximité avec le TDH.

### 3.2.3 Extraction du modèle à partir des mesures

Le modèle de Hammerstein permet une représentation efficace d'un système non-linéaire ainsi qu'un lien facile avec le taux de distorsion harmonique. Cependant, il est nécéssaire de trouver une méthode efficace pour extraire des mesures tout ou une partie de ce modèle. Pour réaliser ceci, nous nous sommes inspirés d'une méthode proposée par Farina *et al.* [Far00].

Cette méthode repose sur une propriété matématique particulière d'une fonction appelée "sweep logarithmique" (cf. équation 3.6). Il s'agit d'un sinus dont la fréquence évolue exponentiellement de $f_1$ à $f_2$ en un temps $T$ (les paramètres $\beta$ et $\tau$ sont exprimables en fonction de $f_1, f_2$ et $T$).

$$x(t) = \alpha \sin(\beta[\exp(\frac{t}{\tau}) - 1]) \quad (\alpha, \beta, \tau) \in \mathbb{R}^3 \tag{3.6}$$

La propriété de ce type de signaux est explicitée par l'équation 3.7 avec $X_n(f) = TF[x^n(t)](f)$ et "TFI" la transformation de Fourier inverse. De façon simplifiée, il est possible de dire qu'un sweep logarithmique élevé à la puissance n est quasiment proportionnel au sweep original, mais en avance de $T_n$ par rapport à ce dernier (cf. figure 3.6). Il est particulièrement intéressant de remarquer que

le temps $T_n$ d'avance de la puissance n-ième sur l'original est une fonction croissante de $n$ et dépend uniquement des paramètres $T, f_1$ et $f_2$.

$$\forall n \in \mathbb{N} \quad TFI[\frac{X_n(f)}{X_1(f)}](t) \simeq \Gamma_n \delta(t - T_n) \quad T_n = T\frac{\ln(n)}{\ln(f_2/f_1)} \quad \Gamma_n \in \mathbb{R} \tag{3.7}$$

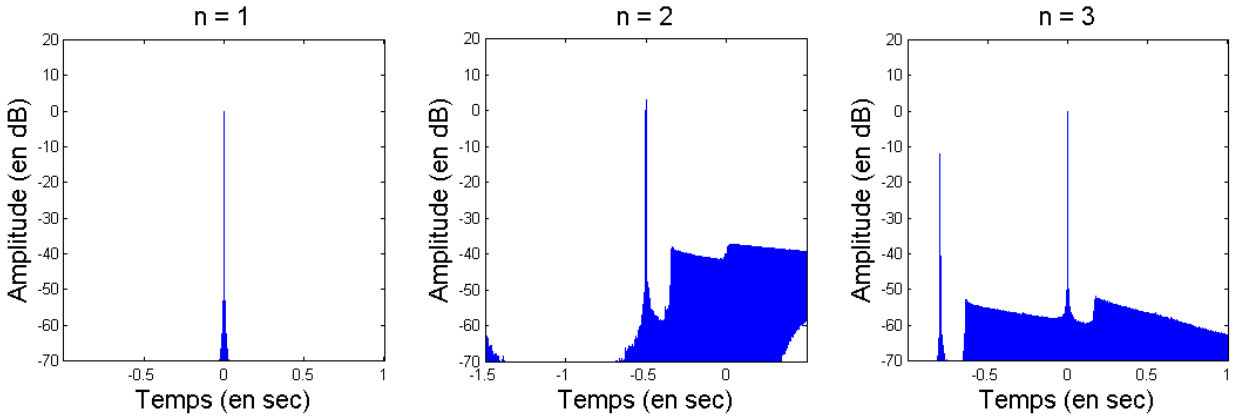



FIG. 3.6 – Vérification numérique de la propriété des sweeps logartithmiques (équation 3.7) par le tracé de $TFI[\frac{X_n(f)}{X_1(f)}](t)$. On constate que les Diracs ressortent bien du bruit de fond et sont bien décalés temporellement les uns des autres. De plus, pour $n = 3$, un deuxième Dirac apparait la ou n'apparait normalement que la contribution pour $n = 1$. C'est un défaut de la méthode qui peut être facilement compensé en soustrayant la composante parasite d'ordre 3 à la composante d'ordre 1.

Si l'on applique à présent la transformation de Fourier inverse (TFI) au rapport de la sortie d'un système représenté par un modèle de Hammerstein (équation 3.2) et dont l'entrée est un sweep logarithmique sur le sweep logarithmique original, on obtient alors l'expression 3.8.

$$TFI[\frac{Y(f)}{X(f)}](t) = \sum_{n=1}^{+\infty} \Gamma_n h_n(t - T_n) \tag{3.8}$$

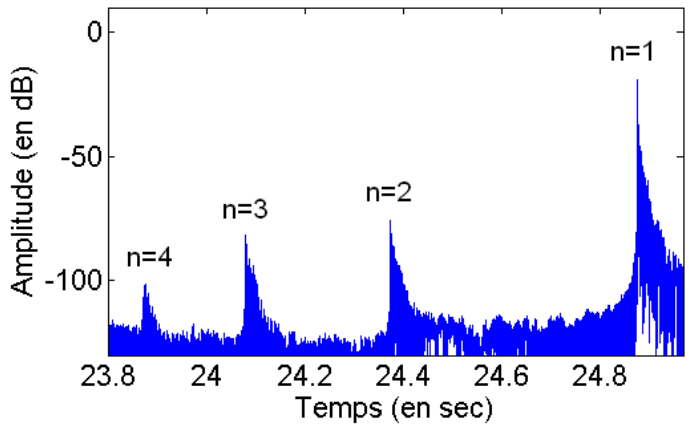



FIG. 3.7 – Allure du résultat de la déconvolution pour des mesures effectuées sur le haut-parleur de référénce, dans l'axe (expression 3.8). Les 4 premières réponses impulsionnelles non-linéaires ressortent bien du bruit de fond.

Alors avec cette méthode de déconvolution, pour peu que les paramètres $T, f_1$ et $f_2$ soient bien choisis, il est possible d'obtenir séparément avec un simple fenêtrage temporel chacune des réponses impulsionnelles non-linéaires $h_n(t)$. Le coefficient $\Gamma_n$ est obtenu de façon numérique en utilisant l'expression 3.7. Il convient de noter que les réponses impulsionnelles non-linéaires $\{h_n(t)\}_{n>1}$ sont situées temporellement avant la réponse impulsionnelle du système $h_1(t)$ et donc ne sont pas noyées dans une éventuelle réverbération (cf. figure 3.7).

Ainsi, avec un seul sweep logarithmique, cette méthode permet d'estimer une représentation complète du système par un modèle de Hammerstein en cascade. Pour des raisons de niveau de bruit nous n'avons pas extrait de réponse fréquentielle non-linéaire $H_n(f)$ d'ordre supérieur à 4 des mesures (voir figure 3.8).

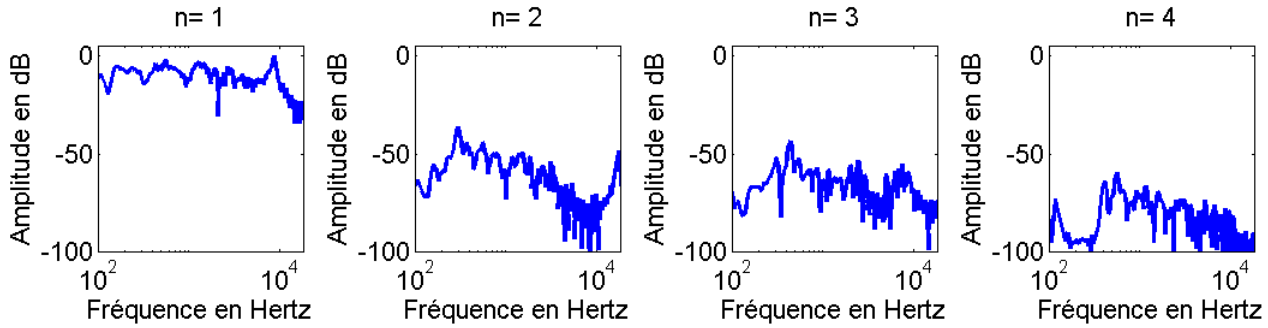



FIG. 3.8 – Tracé de $20\log(|H_n(f)|)$ pour $n \in [1:4]$ dans le cas d'une mesure réalisée dans l'axe du haut-parleur électrodynamique de référence.

## 3.3 Exploitation des mesures

Dans cette section, l'exploitation des modèles de Hammerstein en cascade extraits en chaque point de mesure va être présentée.

### 3.3.1 Mise à l'échelle fréquentielle

Le modèle d'Hammerstein en cascade permet d'estimer de façon équivalente les réponses impulsionnelles ou fréquentielles non-linéaires en chaque point de mesure. Cependant, pour exploiter les mesures, une représentation efficace (i.e. permettant une interprétation facile et intuitive) de ces phénomènes non-linéaires est nécéssaire.

Comme le montre l'équation 3.4, à une fréquence d'excitation $f$, la réponse linéaire est donnée par le produit de la composante de l'entrée par $|H_1(f)|$ assortie d'un déphasage de $\arg(H_1(f))$. La réponse non-linéaire d'ordre $n$ est donné, de façon analogue par le produit de la composante d'entrée par $|H_n(nf)|$, toujours accompagné d'un déphasage de $\arg(H_n(nf))$.

Pour avoir une représentation, dans le domaine fréquentiel, de l'importance relative des différents harmoniques, il est alors nécéssaire de comparer entre eux les gains $\{|H_n(nf)|\}_{n\in\mathbb{N}}$. Nous proposons donc d'introduire les fonctions de comparaisons $\{|C_n(f)|\}_{n\in\mathbb{N}}$ définies par l'équation 3.9 pour avoir une interprétation plus simple des résultats (cf. figure 3.9).

$$\forall n \in \mathbb{N} \quad C_n(f) = H_n(nf) \tag{3.9}$$

Une conséquence logique de ce choix est qu'avec une fréquence d'échantillonnage $f_E$ donnée (ici 44.1 kHz), alors nous n'avons une représentation de $C_n(f)$ que pour $f \in [0 \quad \frac{f_E}{2n}]$.

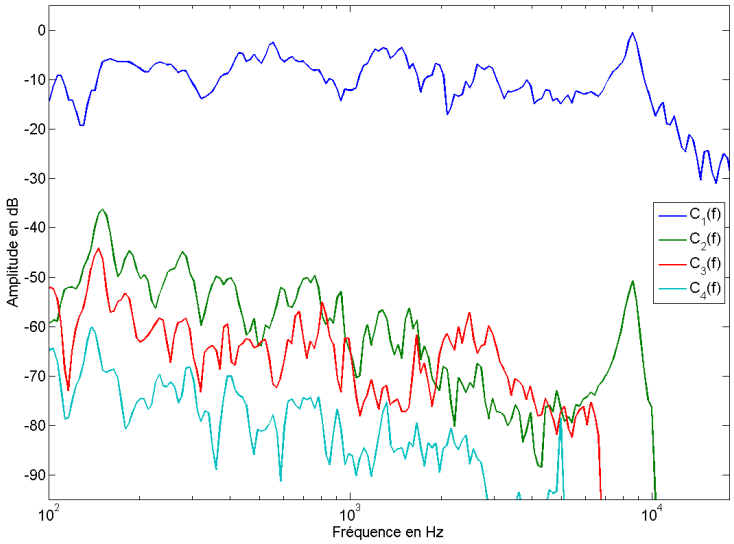



FIG. 3.9 – Représentation de $20\log(|C_n(f)|)$ pour $n \in [1:4]$. Mesures réalisées dans l'axe du haut-parleur utilisé comme référence. Pour une fréquence $f$ donnée, il est possible de lire directement l'amplitude des harmoniques 2, 3 et 4 par simple translation verticale.

### 3.3.2 Ecart au modèle monopolaire

Notre objectif est d'évaluer la qualité d'un système de rendu audio de type MAP dans le cadre d'une restitution pour la WFS. En nous basant sur l'équation 2.2, c'est le modèle monopolaire parfait qui est à priori le système à atteindre. Nous avons donc, dans la représentation spatiale des données, chercher à rapprocher les rayonnements mesurés du rayonnement du modèle monopolaire idéal.

Nous effectuons par soucis de praticité les mesures dans un plan, et non sur une sphère ce qui permettrait un rapprochement aisé avec le modèle monopolaire. Une double compensation est donc mise en oeuvre pour ramener le plan de mesure réel $\Pi$ à une portion de sphère fictive $\mathcal{S}$ de rayon $R$ et qui lui serait équivalente :

- *Une compensation de niveau* : L'amplitude du champ mesuré à une distance $r$ de l'origine est multipliée par $r/R$ pour respecter l'atténuation théorique en $1/r$ et ramener le point de mesure sur la sphère.
- *Une compensation de phase* : Les réponses impulsionnelles linéaires $h_1(t)$ sont toutes alignées temporellement avant le fenêtrage pour l'extraction du modèle (équation 3.8). Alors, les temps de propagation entre l'origine et les points de mesures deviennent tous égaux, ce qui permet de se ramener temporellement au modèle sphérique.

Sur cette portion de sphère fictive $\mathcal{S}$ qui est ainsi créée, la réponse du modèle monopolaire parfait est bien connue : niveau et phase constants pour tous les points de mesure et ce quelle que soit la fréquence. Dans cette optique, les différences de niveau et de phase observés par la suite après les mesures réalisées sur les systèmes réels pourront être interprêtés commes des écarts de rayonnement au modèle monopolaire et nous renseignerons sur la validité d'un tel système pour un rendu de type WFS.

### 3.3.3 Représentations spatiales des composantes linéaires et non-linéaires des fonctions de comparaisons

Nous sommes maintenant en mesure de représenter et d'interprêter en chaque point les fonctions $\{|C_n(f)|\}_{n\in\mathbb{N}}$, qui sont en quelque sorte des images du champ de pression acoustique rayonné par le système étudié pour chaque fréquence. Jusqu'ici la dépendance spatiale de ces fonctions n'a jamais

été explicitement mentionnée. Nous les noterons donc à présent $\{|C_n(M,f)|\}_{n\in\mathbb{N}}$ ou $M$ est le point de mesure ramené sur la sphère fictive pour ne pas oublier ce point.

Nous avons alors choisis trois représentations spatiales de $\{|C_n(M,f)|\}_{n\in\mathbb{N}}$ pour dresser des cartes d'écarts au modèle monopolaire (cf. figure 3.10) :

1. *Représentation sphérique* : C'est une représentation classique en trois dimensions, ou les données sont représentées en coordonnées sphériques. Un rayon $r$ proportionnel à l'amplitude mesurée en dB est attribué dans la direction $(\theta,\phi)$ ou il a été mesuré.
2. *Carte de niveau sans phase* : Il s'agit d'une représentation plane ou l'amplitude du champ en dB est représentée par des courbes de niveau et une grille de couleurs dans le plan $(\theta,\phi)$.
3. *Carte de niveau avec phase* : C'est une représentation similaire à la précédente sauf que les couleurs chaudes indiquent une phase entre 0 et $\pi$ tandis que les couleurs froides indiquent une phase entre $\pi$ et $2\pi$. De plus l'amplitude mesurée n'est pas ici exprimée en dB.

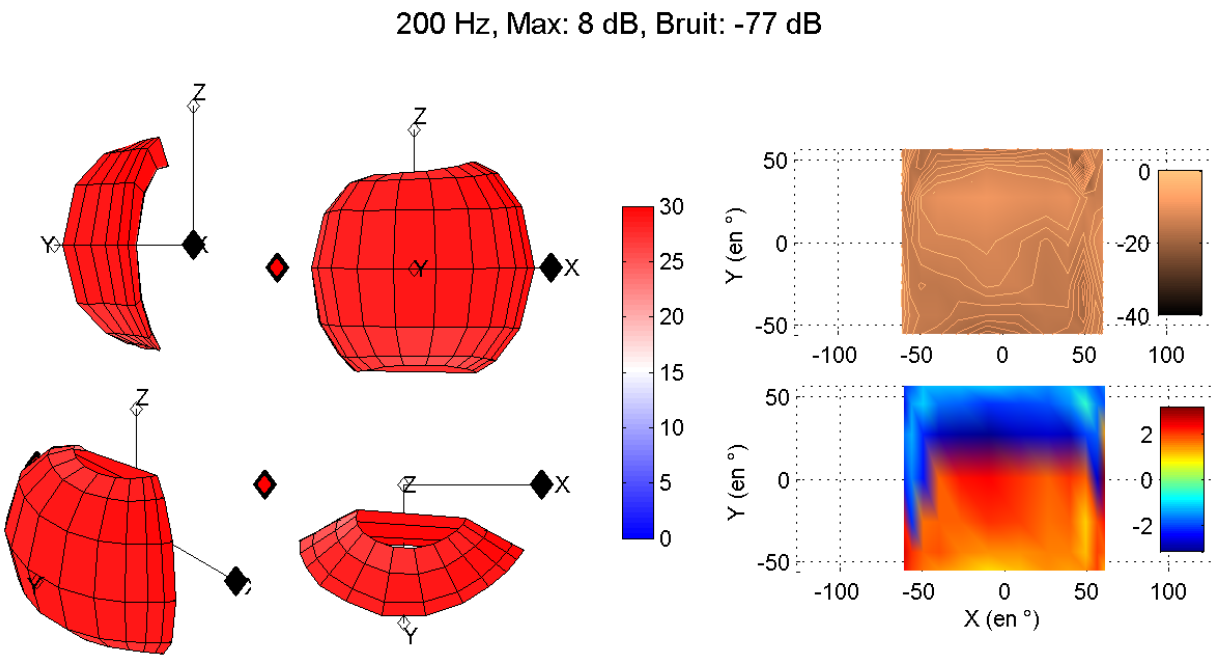



FIG. 3.10 – Les représentations 1, 2 et 3 sont illustrées ici, pour la partie linéaire du rayonnement du haut-parleur utilisé comme référence, à la fréquence de 200 Hz. A gauche, représentation 1, en haut à droite représentation 2 et en bas à droite représentation 3.

Les figures correspondant au réponses fréquentielles dans l'axe et aux écarts au modèle monopolaire, dans les trois représentation, pour la partie linéaire et pour le TDH sont données dans l'annexe A pour chacuns des systèmes étudiés.

### 3.3.4 Définition d'indices compacts

Les représentations spatiales présentées précédement permettent d'étudier en détail le champ acoustique rayonné par le système étudié, mais ne permettent pas de comparaison facile des systèmes entre eux. Nous proposons donc deux catégories d'indices compacts qui permettrons de juger de la qualité de restitution offerte par les systèmes étudiés :

– Indices liés à la qualité de la composante linéaire du rayonnement.
– Indices liés à l'importance de la composante non-linéaire dans le rayonnement.

Ces indices dépendent de la fréquence et ont été calculés sur des bandes de tiers d'octave pour être en accord avec la résolution de l'oreille humaine [Bla97].

Il est important de noter que les indices définis ci-dessous n'ont d'interprétation possible que dans le contexte défini précédement et ne sont pas représentatif dans l'absolu du système, mais sont la comme moyen de comparer entre eux les systèmes.

#### Qualité de la composante linéaire

Les deux qualités attendues d'un haut-parleur, en terme de composante linéaire et conformément à une utilisation en WFS sont :

1. Une réponse fréquentielle plate et ce dans toutes les directions.
2. Un rayonnement homogène dans l'espace.

Pour quantifier le permier point, nous avons choisi d'évaluer la densité spectrale de pluissance émise par les systèmes, au travers de $\mathcal{S}$, sur des bandes de fréquence en tiers d'octave. La définition de cet indice, nommé $DSP_{lin}(f)$, est donnée par l'équation 3.10. Cet indice nous renseigne donc sur la capacité du système à rayonner, globalement dans l'espace, de la puissance dans une bande de fréquence donnée.

$$DSP_{lin}(f) = \iint_{\mathcal{S}} |C_1(M,f)|^2 dS \tag{3.10}$$

Pour quantifier le second point, nous avons choisi d'évaluer l'écart type de la répartition statistique composée des écarts en dB au modèle monopolaire sur $\mathcal{S}$, par bande de tiers d'octave. La définition de cet indice, nommé $STD_{lin}$ est donnée par l'équation 3.11, ou *std* est l'écart type. Cet indice donne donc des informations sur la variabilité spatiale de la fonction $|C_1(M,f)|$ sur $\mathcal{S}$.

$$STD_{lin}(f) = std(\{20\log(|C_1(M,f)|)\}_{M\in\mathcal{S}}) \tag{3.11}$$

Ces deux indices vont donc nous servir à évaluer pour un système donné la qualité de son rayonnement.

#### Importance de la composante non-linéaire

En ce qui concerne la partie non-linéaire du rayonnement, il est surtout important de s'assurer que la part de la composante non-linéaire, relativement à la part de la composante linéaire n'est jamais trop importante.

Pour quantifier ce point, nous avons donc choisi d'évaluer le rapport de la densité spectrale de puissance émise par les composantes non-linéaires sur la densité spectrale de puissance émise par la composante linéaire, à travers $\mathcal{S}$, pour chacune des bandes de tiers d'octave. La définition de cet indice, appelé $DSP_{dist}(f)$ est donnée par l'équation 3.12.

$$DSP_{dist}(f) = \iint_{\mathcal{S}} \frac{|C_2(M,f)|^2 + |C_3(M,f)|^2 + |C_4(M,f)|^2}{|C_1(M,f)|^2} dS \tag{3.12}$$

Cet indice, plus ésotérique trouve sa justification dans le fait que la réponse fréquentielle des MAP n'est pas plate (cf. l'annexe A). Par conséquent, le TDH est très variable dans l'espace , car fortement corrélé aux modes des panneaux. L'intégration de ce taux de distorsion harmonique sur la surface de mesure (ce qu'est quasiment $DSP_{dist}$) permet donc d'avoir un indice qui soit plus porteur de sens que le TDH dans l'axe.

## 3.4 Résultats

Dans cette partie, les résultats obtenus sur les différents systèmes dans le contexte précédements sont présentés ainsi que les conclusions que l'on peut en tirer par rapport aux objectifs fixés.

### 3.4.1 Influence de l'encastrement

Le premier objectif était d'étudier si l'encastrement d'un MAP avait une influence sur son rayonnement, et si tel était le cas, quel était alors le meilleur type d'encastrement pour un rendu de type WFS. Les indices compacts $DSP_{lin}(f)$ et $STD_{lin}(f)$ ont donc été calculés à partir des mesures réalisées sur chacun des systèmes et sont représentés sur la figure 3.11.

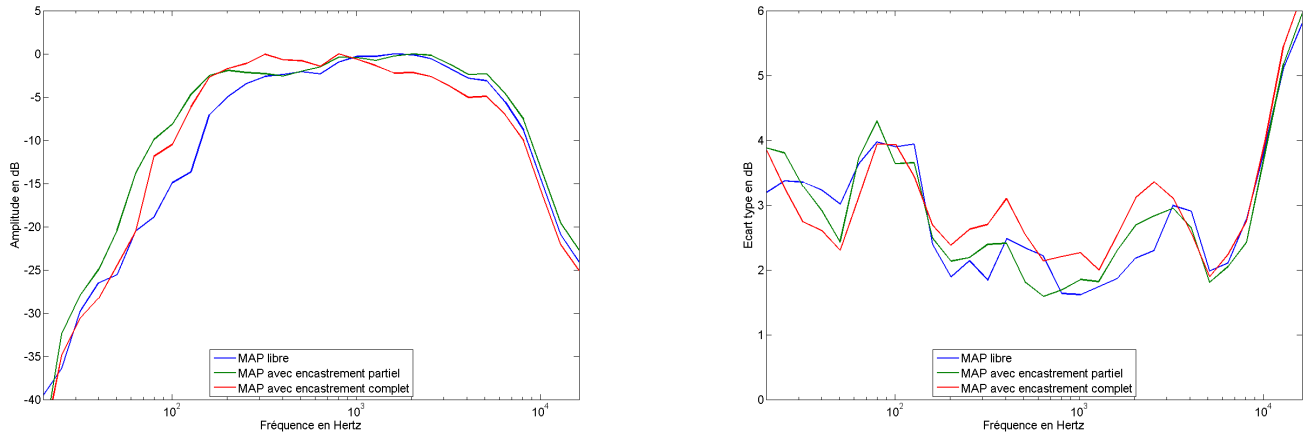



FIG. 3.11 – Comparaison des indices compacts $DSP_{lin}(f)$ (à gauche) et $STD_{lin}(f)$ (à droite) pour les 3 types d'encastrements.

Du point de vue de la qualité du rayonnement dans l'espace, on peut constater en observant l'indice $DSP_{lin}(f)$ que d'une manière générale, les 3 MAP's ont une densité spectrale homogène sur une bande de fréquence allant de 100 Hz à 10 kHz. On observe aussi que le MAP libre rayonne moins bien en basse fréquences que les autres. Cela peut être dû au fait que les premiers modes rayonnants, pour les deux types d'encastrement extrémaux (libre et encastrement complet), qui sont situés dans la même région fréquentielle sont très dissipatifs et sujets à des court-circuits acoustiques. L'encastrement complet, en ayant tendance à rigidifier l'ensemble de la structure, pourrait alors expliquer un meilleur rendement en basses fréquences. Le phénomènes inverse est observé en hautes fréquences. Cette fois-ci, c'est le MAP avec encastrement total qui rayonne le moins bien. Cela peut être corrélé avec le fait que nous ayons disposé sur l'ensemble de son pourtour une mousse absorbante pour le maintenir. L'action de cette mousse pourrait être d'agir comme un système amortisseur en hautes fréquences, dissipant alors une partie de l'énergie qui était auparavent rayonnée.

L'analyse de l'indice $STD_{lin}(f)$ montre une variabilité moyenne de la densité spectrale de puissance quie se situe autour de 2 dB dans la zone fréquentielle ou les MAP's sont efficaces (de 100 Hz à 10 kHz). Cette variabilité, bien que perceptible [Bla97] n'est cependant pas très importante. De plus, cette figure nous fait aussi dire que la variabilité dans le rayonnement du MAP avec encastrement total est globalement plus importante sur l'ensemble des fréquences audibles. Cela aurait donc tendance à confirmer l'hypothèse intuitive selon laquelle l'encastrement total rigidifierait les modes. Entre le MAP encastré partiellement et le MAP libre, on ne note pas de différence significative.

Le MAP avec encastrement partiel semble donc être un plutôt bon compromis pour avoir une densité spectrale de puissance assez homogène fréquentiellement et spatialement. C'est donc cette solution qui sera retenue pour la réalisation du MAP de grandes dimensions.

### 3.4.2 Non-linéarités rayonnées par les systèmes

Un autre de nos objectifs était l'étude des non-linéarités rayonnées par ce type de structures. Sur la figure 3.12, l'indice compact $DSP_{dist}(f)$, correspondant à cet objectif est représenté en fonction de la fréquence.

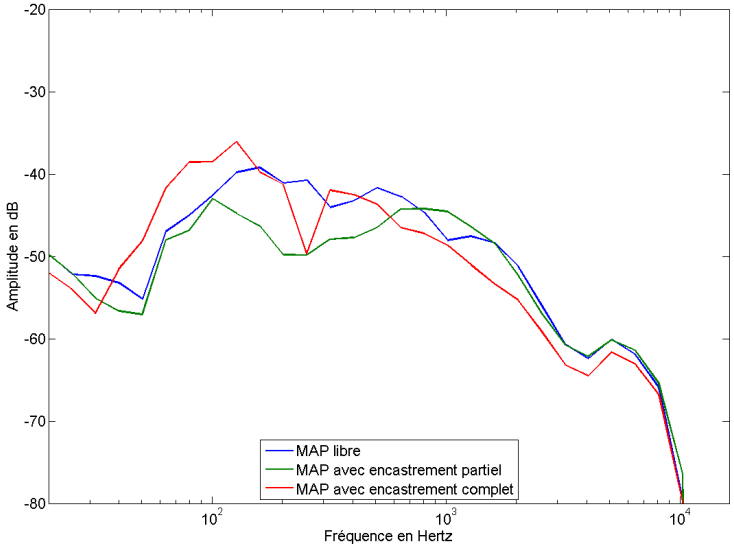



FIG. 3.12 – Comparaison de l'indice compact $DSP_{dist}(f)$ pour les 3 types d'encastrement.

Un commentaire général à propos de cette figure est que sur la bande de fréquence ou les MAP rayonnent bien de façon linéaire (i.e. entre de 100 Hz à 10 kHz) alors la densité spectrale de puissance de la distorsion harmonique est située aux alentours de -40 dB, les mesures ayant été effectuées avec un niveau sonore de 97 dB. La distorsion harmonique, de l'ordre du % à 97 dB n'est donc pas très importante et de ce point de vue les MAP's peuvent être jugés comme étant de bonne qualité. Toutefois, cette affirmation est à nuancer car, comme spécifié dans l'introduction de la section 3.3.4, la validité de ces indices dans un contexte général n'est pas évidente et la concordance avec des mesures standard de TDH (même si un lien mathématique existe) n'a pas été vérifiée. Il s'agit, pour nous, surtout d'un outil de comparaison entre les systèmes.

Les trois dispositifs n'ont cependant pas tout à fait la même contribution non-linéaire. En effet, il semble globalement que les MAP libre et avec encastrement total produisent plus de non-linéarités que le MAP avec encastrement partiel (environ 5 dB). En premier lieu, un résultat qui semble surpenant est d'obtenir des non-linéarités présentes sur les bords du MAP libre. Une hypothèse, à vérifier, est que ces non-linéarités seraient dûes à la création de flux aerodynamique au niveau des arêtes du MAP, qui sont très anguleuses. En ce qui concerne le MAP avec encastrement total, la densité spectrale du rayonnement non-linéaire est plus importante en basses fréquence qu'en haute fréquence. Pour la partie basse fréquence, cela peut être corrélé avec la rigidification des modes, qui induit de plus grand déplacement et donc est suceptible de faire apparaître plus de non-linéarités géométriques que le MAP libre (faisant intervenir un modèle de Von Karman par exemple [Tou08]). D'autre part, l'atténuation observée en hautes fréquences, peut la encore s'expliquer par l'action de la mousse qui jouerait le rôle d'amortisseur sur les côtés augmentant ainsi les pertes d'énergies.

Le MAP avec encastrement partiel présente, comme dans la sous-section précédente, un comportement intermédiaire entre ces deux types d'encastrements extrême qui s'avère interéssant. Cela nous conforte donc dans le choix que nous avons fait.

### 3.4.3 Validité du MAP avec encastrement partiel par rapport au haut-parleur de référence

Le choix que nous avons donc réalisé parmi les différents systèmes à l'étude a été le MAP avec encastrement partiel. Cependant, pour conclure cette étude, il peut être interéssant de voir dans quelle mesure, à la lumière des indices compacts créés pour l'occasion, ce MAP diffère du haut-parleur de

référence. Les trois indices compacts sont comparés pour les deux systèmes à la figure 3.13.

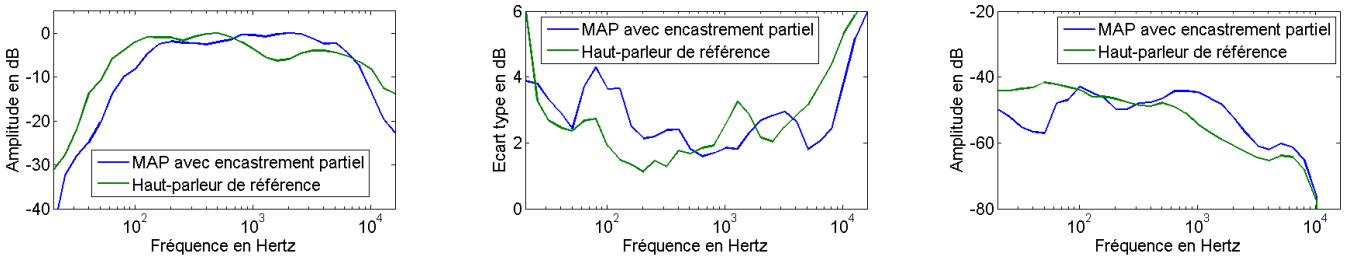



FIG. 3.13 – Comparaison des indices compacts $DSP_{lin}(f)$ (à gauche), $STD_{lin}(f)$ (au milieu) et $DSP_{dist}(f)$ (à droite) pour le MAP encastré partiellement et le haut-parleur de référence.

Tout d'abord, l'étude de $DSP_{lin}(f)$ révèle que le haut-parleur électrodynamique présente un meilleur comportement en basse fréquences que le MAP étudié. Ceci semble cohérent, compte tenu de la présence de possibles court-circuit acoustiques dans le cas du MAP et des limitations induites par l'excitateur utilisé. Par contre, en hautes fréquences, le MAP semble présenter une densité spectrale de puissance plus homogène que le haut-parleur. Cela peut être rapproché du fait, déja constaté dans [Cor07], que les MAP sont moins directifs en hautes fréquences que ne le sont les haut-parleurs électrodynamiques. Effectivement, c'est le cas, et donc lors de l'intégration sur la surface de mesure équivalente, pour le calcul de l'indice, la puissance rayonnée est moindre pour le haut-parleur.

Ensuite, l'analyse de la courbe du milieu, donnant l'évolution de $STD_{lin}(f)$ avec la fréquence permet un commentaire intéresssant. En effet, on constate que la variabilité du rayonnement émis par le MAP avec encastrement partiel est plus importante que celle du haut-parleur électrodynamique. Cela est dû au comportement modal des MAP's qui est inhérent à leur structure. Il est par contre envisageable qu'un MAP de plus grande dimension, et donc présentant une plus grande densité modale que le MAP ici présent, voit la variabilité de sa densité spectrale de puissance diminuer.

Enfin, l'indice $DSP_{dist}(f)$, ne permet pas de dire beaucoup sur les différences entre les deux systèmes. Globalement ils semblent équivalent à ce niveau sonore, même si le haut-parleur électrodynamique présente plus de distortion en basse fréquence.

Le choix du MAP avec encastrement partiel parait donc réaliste pour une utilisation en WFS et présente des caractéristiques quand même assez proches de celles du haut-parleur électrodynamique de référence.

## Conclusion

Dans ce chapitre une étude acoustique ayant pour objectif de caractériser, en fonction de leur encastrement le rayonnement de haut-parleurs de type MAP a été conduite. L'accent a par ailleurs été mis sur les non-linéarités émises par les sytèmes à l'étude et une méthode efficace de modélisation et de mesure a été présentée. Pour pouvoir comparer les systèmes, des indices compacts et signifiant par rapport aux objectifs fixés ont été suggérés et utilisés.

Quelques points restent néanmoins encore délicats. En effet, la validité de la modélisation par rapport à des mesures réelles de TDH n'a pas été vérifiée et donc ne permet pas de comparer les résultats obtenus en terme de TDH avec ceux déja connus pour d'autres types de haut-parleurs. Cependant, un lien mathématique assez simple existe, ce qui suggère que cette vérification est aisément réalisable.

A la lumière de ces indices, il semble donc que le MAP avec un encastrement partiel soit le plus intéressant des trois systèmes de MAP proposés. De plus, comparé au haut-parleur électro-dynamique ses performances sont correctes. C'est donc cette stratégie d'encastrement que nous avons choisis d'adopter pour la réalisation du MAP de grande dimension destiné au SMART-$I^2$.

# Chapitre 4

# Conception, Réalisation et Évaluation du SMART-$I^2$

Ce chapitre présente la façon dont le SMART-$I^2$ a été conçu et les différents choix technologiques réalisés. Puis, il aborde l'évaluation de la qualité de localisation proposée par ce dispositif.

L'intégralité de ce chapitre fait l'objet d'un article accepté pour être publié dans les Proceedings de la 125 ème Convention de l'Audio engineering Society, et est par conséquent rédigé en anglais. L'analyse objective a été réalisée par Etienne Corteel.

## Introduction

In recent years, the advancement of immersive environments has separately produced systems with improved quality for 3D stereoscopic graphical rendering and also for 3D audio rendering. Despite these advances, few combined modality systems of high quality have been developed. This difficulty can be attributed to the different and stringent technical requirements for each rendering system, primarily in terms of equipment (construction material and placement). As so, these devices actually provide the users only a limited presence sensation.

In this paper, presence is understood as *"the perceptual illusion of non-mediation"* [Lom97]. Many dimensions are involved in this high level cognitive process which includes communication and interaction. From a technological point of view, perceptual realism conveyed by the rendering interface is often thought of as a way to increase presence independently of the task or the content. The rendering device should not be detected by the user. It should be conceived as a large open window through which the users experience the virtual world. Users should have the impression of sharing the same space and should be able to interact together or within the environment. These aspects have lead to the conception of the SMART-$I^2$.

In devices where a global audio rendering is chosen, rather than headphones (in CAVE-like environments for example [Cru92][EVE], or frontal screen environments [Bru04][Spr06]), the optical pathways between the projector, screen, and the users eyes are often in conflict with the optimal acoustic pathways between loudspeakers and the users ears. So, some architectural and physical compromises must often be made in order to provide an effective audio-visual rendering, and loudspeakers are often placed behind or on the edges of the large rear-projection screen thereby reducing audio quality for the sake of graphical quality. The acoustical transmission characteristics of the screen and non-ideal positioning make it very difficult to achieve fine control of the acoustic field in the rendering zone. This is problematic for high quality precise advanced spatial rendering technologies such as Ambisonics [Ger73], high order Ambisonics (HOA) [Dan00], vector based amplitude panning (VBAP) [Pul97] or wave field synthesis (WFS) [Ber93]. Some solutions have been considered to compensate for the transmission characteristics effect of rear projection screens [Lok07] but a fine acoustic control is still very difficult as loudspeaker placement is still highly guided by the optical pathways. Therefore, in such devices, perceptual realism is not maximum.

Another common way to reproduce spatial audio is binaural rendering over headphones using the head-related transfer function (HRTF).[Beg94] The quality of binaural rendering is highly linked to the use of an individual HRTF rather than a non-individual HRTF. In addition, head tracking and a dedicated rendering engine are usually necessary for each user. The use of headphones can reduce the level of immersion and also communication and interactivity for multi-user systems. Thus, loudspeaker reproduction has certain advantages over headphones, and is the option chosen in this system.

This research study presents a novel design concept, the SMART-$I^2$, which attempts to resolve the problem of high quality immersive audio-visual rendering by rendering coherent audio and visual information. The sound rendering is realized using wave field synthesis technology (WFS) [Ber93][Cor06ME] which relies on physical based reproduction of sound fields within an extended listening area. The graphical scene is rendered using tracked stereoscopy [Fro05], which presents users with the correct rendered visual image for each eye separately. Finally, these two technologies are combined together using two large Multi-Actuator Panels [Boo04][Cor07] which act both as projection screens and as loudspeaker arrays.

In this paper, the SMART-$I^2$ architecture is first described and technological choices are explained. The spatial audio-visual rendering of the SMART-$I^2$ is then assessed with a combined objective and subjective evaluation. The validation focuses on localization accuracy for both static and moving users.

## 4.1 The SMART-$I^2$ system

The SMART-$I^2$ system is an advanced audio-visual interface which results from the scientific collaboration of LIMSI-CNRS and sonic emotion. In this part, the key ideas of the system are explained and the realized system is later presented.

### 4.1.1 Audio-visual consistency over a large area

The key concept of the SMART-$I^2$ is to create a virtual window through which a plausible virtual world is perceived. All the spatial properties of the audio-visual scene should be accurately conveyed to the user(s) at any point within the rendering area. This includes angular and distance perception which should remain accurate throughout. The audio-visual window therefore ensures that static but also dynamic localization cues, such as the motion parallax effect, are preserved. The motion parallax effect occurs when users are moving about the rendering area. This effect is linked to the user's ability to ascertain the positions of audio-visual objects using the relative movements of these objects. For example, in the scene presented in figure 4.1, the user can ascertain the relative distance between objects 1 and 2 by comparing all the distance and angular information coming from these objects when moving from point A to point B.

WFS rendering, using horizontal loudspeaker arrays, is often regarded as an acoustical window which is limited by the array extension. For a linear array, the audio rendering is therefore limited to the horizontal plane but remains valid within a large listening area which can be determined using visibility criteria of the source through the loudspeaker array. Unlike other loudspeaker based spatial rendering techniques, WFS naturally gives access to the "audio" motion parallax effect.

3D visual rendering requires one to independently address each eye of the user. The rendering should also be adapted to the position and orientation of the user in order to render always the correct point of view and to preserve the visual scene organisation (objects sizes and angular positions). This technique is referred to as tracked visual stereoscopy. It combines visual crosstalk cancellation using light polarization ("passive" stereoscopy) or time multiplexing ("active" stereoscopy) with the adaptation of graphic rendering to the current position of the user. The user should wear special glasses for visual crosstalk cancellation. These are special polarized glasses (different polarization for each eye and each projector) for passive stereoscopy and shutter glasses synchronised with the video rendering for active stereoscopy.

Active stereoscopy is the most efficient crosstlak cancellation technique. However, it is expensive since it requires electronic synchronization of the projectors and the glasses. It is also known to induce

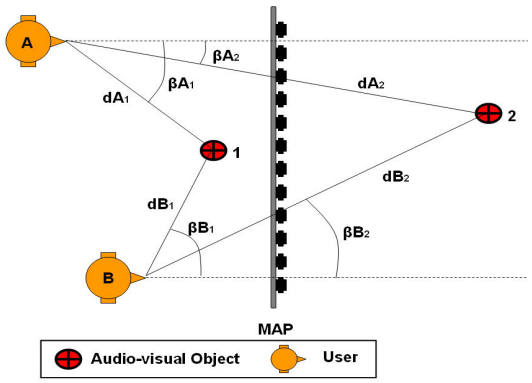



FIG. 4.1 – Illustration of the motion parallax effect.

visual fatigue which can be disturbing in the long run. Therefore, it was chosen to rely on “passive" stereoscopy. Both techniques could also be combined to increase the number of users. This is a topic for further research.

WFS and tracked stereoscopy can both be thought as a perceptual window opening onto a virtual world. They are thus coherent technological choices.

### 4.1.2 Overview of the system

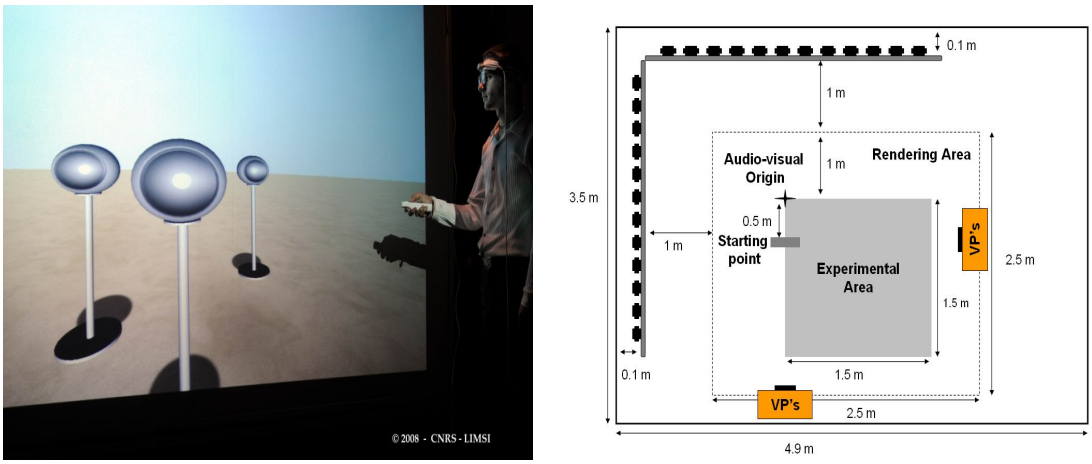



FIG. 4.2 – Overview of the SMART-$I^2$ on the left. Schematic view of the SMART-$I^2$ installation on the right.

The SMART-$I^2$ project was conceived as a large audio-visual window where the screens would also act as loudspeakers using WFS rendering. The installation of the system at LIMSI-CNRS is shown in figure 4.2. Two large screens form a corner so that the system presents a wide angular opening and a large listening area (2.5 m $\times$ 2.5 m). This corner installation poses less constraints, avoiding many of the common practical and acoustical concerns of a partly closed box as used in large CAVE systems. The walls and ceiling of the current installation are mostly covered with 2” acoustsics foam to limit reflections. For the current validation study, only one screen is mounted (see figure 4.2).

### 4.1.3 Large MAP as a projection screen and loudspeakers array

MAP loudspeakers are derived from DML technologies. The panels are constructed of a light and stiff material on which multiple actuators are attached. Each actuator is independently driven. A

multichannel loudspeaker is thus realized with a single unique physical structure.

Thanks to their low visual profile, multiple MAP loudspeakers can provide seamless integration of tens to hundreds of loudspeakers in an existing environment. Due to the excitation nature of the panel, and their generally large size, displacement distances are very small and do not disturb 2D or even 3D video projection on the surface of the panel.

### Design of the panels

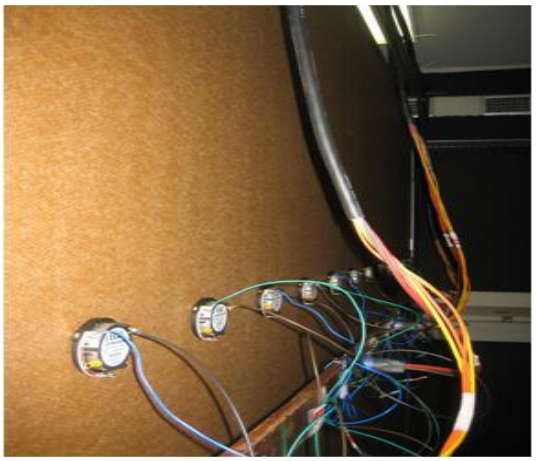


FIG. 4.3 – View of the back of the SMART-$I^2$ MAP.

The MAP used for the SMART-$I^2$ is a large honeycomb panel, 2 m × 2.66 m. The dimensions of the panel present a classical 4/3 ratio between height and width suitable for video projection. Twelve exciters are attached to the rear of the panel. They are located on a horizontal line every 20 cm. Such spacing corresponds to an aliasing frequency of about 1500 Hz accounting for the loudspeaker array size and the extension of the listening area.[Cor06AL] This aliasing frequency assures an efficient spatial audio rendering. Some strategies are available to further raise the aliasing frequency.[Cor08SW]

The front face of the panel has been treated with a metallic paint designed to maintain light polarization thus allowing for the use of passive stereoscopic projection. The panel is mounted in an aluminum frame such that the exciter array is at a height of 1.5 m (see figure 4.3). This positioning ensures that the horizontal plane where audio reproduction is correct corresponds to typical ear height for the majority of users.

### Validity of the MAP for WFS

It has been shown that MAPs can be used for WFS rendering.[Boo04][Cor07] However, panels considered in previous studies were significantly smaller (40 cm × 60 cm or 133 cm × 67 cm) than those used here. Pueo *et al.* recently studied larger panels [Pue08], showing measurements of reproduced sound fields using a 2.44 m wide and 1.44 m high panel. These objective results tend to confirm the potential use of such large panels for WFS, which should nevertheless be confirmed by further perceptual studies.

The typical frequency response of a MAP is not as optimized as electrodynamic loudspeakers, and a large MAP is not expected to have the same radiation as smaller scale MAP loudspeakers. In figure 4.4, the frequency response of the MAP designed for the SMART-$I^2$ is given for a central exciter, measured at 60 cm in front of the panel on-axis and at 30°. It can be seen the global frequency response is reasonably flat (±5 dB between 100 Hz and 15 kHz). The frequency responses measured on axis and at 30° off axis are similar. The level is somewhat reduced below 750 Hz by about 3 dB at 30° as compared to on-axis. A full study, including directivity patterns, is a topic for future research.

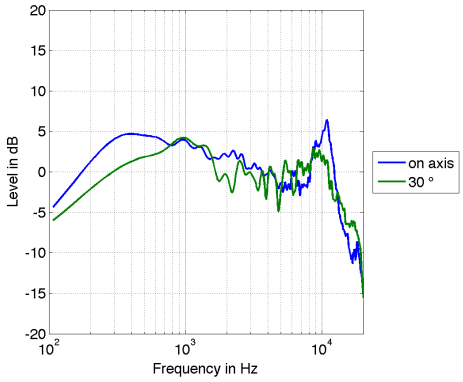



FIG. 4.4 – Frequency response of the large MAP used for SMART-$I^2$.

The frequency response of the large MAP seems appropriate for WFS rendering. In this study, only a naive equalization method is used which compensates for the global radiated response of the loudspeaker array (individual equalization, [Cor06ME]). Advanced multichannel equlization techniques will be employed in further studies.

### 4.1.4 Rendering architecture of the system

The architecture of the SMART-$I^2$ is composed of three rendering components :

- *Virtual Choreographer (VirChor)* : An open source real-time 3D graphics engine that relies on an XML-based definition of 3D scenes with graphic and sonic components.[Jac04][VIR]
- *The WFS Engine* : A real-time audio engine dedicated to low latency (less than 6 ms) WFS rendering which realizes the real-time filtering of up to 24 input channels (virtual sources) to 24 output channels (every exciter of the two MAPs).
- *Max/MSP* : A real-time audio analysis/synthesis engine using a graphical programming environment that provides user interface, timing, and communications.[MAX]

The audio visual scene description is based on the "Scene Modeler" architecture [Bou08], which is a generic package for virtual audio-visual interactive real-time environment creation using both Max/MSP and VirChor. The Model-View-Controller organisation of the system [Bur87] is depicted in figure 4.5.

For real-time rendering, two separated computers are used. On the main machine, Max/MSP and VirChor are running simultaneously. Virchor manages the virtual scene and graphical rendering. Max/MSP receives scene information from VirChor and is responsible for generating the audio content of the scene. Audio channels are sent to the WFS engine which creates the driving signals for the exciters.

All rendering components are connected on a network which transmits the scene description (position of audio-visual objects, position/orientation of users provided by tracking devices, interaction parameters, basic services : start, stop rendering, shutdown machines...). This information is used to perform real-time updates of both the audio and graphic rendering, making the system fully interactive.

## 4.2 Evaluation of the SMART-$I^2$ spatial rendering

In this section, we focus on the evaluation of the spatial rendering quality of the SMART-$I^2$ interface. The evaluation combines an objective analysis of the binaural signals at the user's ear and a subjective analysis based on listening tests.

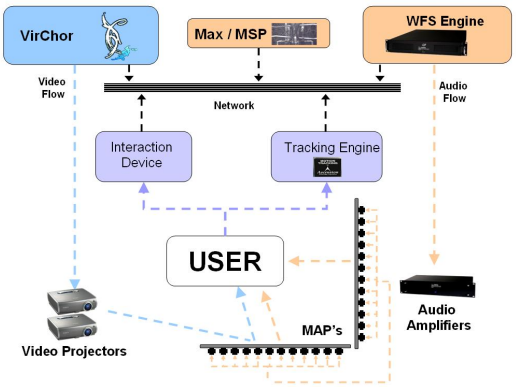



FIG. 4.5 – Model-View-Controller organisation of the SMART-$I^2$.

### 4.2.1 Presentation of objective analysis

The objective analysis presented here is similar to an earlier work by Sanson, Corteel and Warusfel.[Cor08OA] Sanson *et al.* proposed a combined objective and subjective analysis of localization accuracy in WFS using individual HRTFs to binaurally render simulated virtual WFS configurations over headphones.

#### Method

The method used here consists in estimating the localization accuracy of a given loudspeaker based spatial rendering technique using measured or simulated binaural impulse responses.

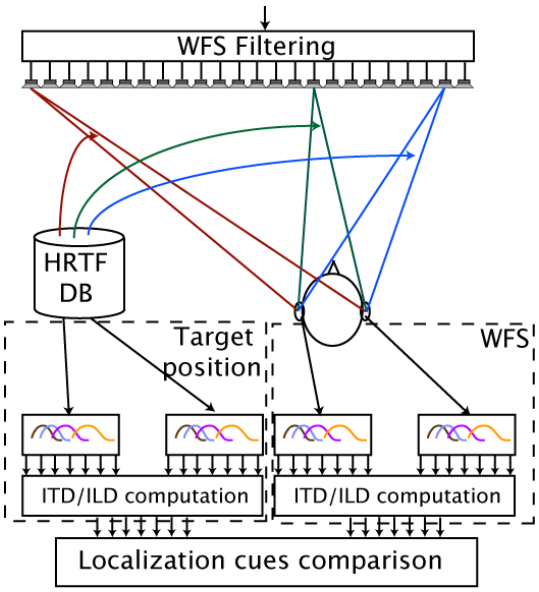



FIG. 4.6 – Extraction of objective localization cues.

This method is composed of five steps illustrated in figure 4.6 :

1. Binaural measurement of a dummy head in an anechoic chamber for a large number of positions (typically every 5° in both azimuth and elevation).
2. Binaural impulse responses on site measurement or estimation (from anechoic binaural measurements [Cor08OA]) at given listening positions.
3. Estimation of “synthesized" binaural impulse responses for a given virtual source by summing the respective contribution of each array transducer.

4. Computation of localization cues (Interaural Time Differences (ITD) and Interaural Level Differences (ILD)) for both "ideal" and "synthesized" binaural impulse responses in auditory bands (40 ERB bands between 100 Hz and 20 kHz).[1]
5. Computation of ITD error, $ITD_{err}(f_c(n))$, and ILD error, $ILD_{err}(f_c(n))$, in each ERB frequency band where $n$ ($f_c(n)$ is the center frequency of the ERB band $n$).

The "ideal" impulse response is extracted from the database of anechoic impulse responses according to the virtual source and the listening position.

Global frequency independent localization accuracy criterion can finally be extracted using the weighting function $q(f)$ proposed by Stern *et al.* in [Ste88] that accounts for the relative importance of frequency bands to localization. Here a normalized weighting function, $q_{norm}(f)$, is defined as :

$$q_{norm}(f) = \frac{q(f)}{mean_{n=1..40}(q(f_c(n)))}. \tag{4.1}$$

The global ITD bias, $B_{ITD}$, is then defined as :

$$B_{ITD} = mean\left[ITD_{err}(f_c(n))q_{norm}(f_c(n))\right], \tag{4.2}$$

and the ITD variability, $V_{ITD}$, as :

$$V_{ITD} = std\left[ITD_{err}(f_c(n))q_{norm}(f_c(n))\right], \tag{4.3}$$

where *std* is the standard deviation. Similarly, an ILD bias, $B_{ILD}$, and ILD variability, $V_{ILD}$, are defined as :

$$B_{ILD} = mean\left[ILD_{err}(f_c(n))q_{norm}(f_c(n))\right], \tag{4.4}$$

and

$$V_{ILD} = std\left[ILD_{err}(f_c(n))q_{norm}(f_c(n))\right]. \tag{4.5}$$

These criteria are given here for wide band stimuli. In cases where the stimuli has only energy in a limited number of frequency bands, the criteria can easily be adapted to account only for these as proposed in [Cor08OA].[2]

### Interpretation of the objective localization criteria

The bias measures $B_{ITD}$ and $B_{ILD}$ indicate the average offset of the conveyed ITD and ILD respectively, compared to the reference (anechoic) ITD and ILD respectively, at the target position. It might be interpreted as an indication of a potential perceptual shift in azimuth due to distorted localization cues conveyed by the considered reproduction setup. The suggested shift would be to the right for positive values or to the left for negative values. The variability measures $V_{ITD}$ and $V_{ILD}$ are an indication of the consistency of the localization cues among the frequency bands. It might thus be considered as an indication of the locatedness of the synthesized virtual source.

In [Cor08OA], Sanson *et al.* show that the abolute localization bias for WFS does not typically exceed 5° independently of the loudspeaker spacing (15 cm or 30 cm in this study) considering large band stimuli covering the entire audible frequency range (white noise) and virtual sources located behind the loudspeaker array. The absolute localization bias typically increased from 7° to 10° for high-passed noises with cutoff frequencies of 750 Hz, 1500 Hz, and 3000 Hz. The associated analysis showed that, in some cases, the ITD and ILD errors where indicating conflicting directions which could be attributed to the finite length of the loudspeaker array. Then, in most of these cases, the ILD cues were dominating which may be due to the frequency content of the employed stimuli (high-passed white noise) and possibly the large variability of the ITD (typically above 0.3 ms). Only in cases where

[1] For a complete description of the ITD and the ILD extraction process, please refer to [Cor08OA]

[2] The criteria presented here intend to provide a more consistent naming convention than that used in [Cor08OA]. They are however very similar in their definition to the $Mean_{err}$ and $Std_{err}$ criteria defined for ITD and ILD in [Cor08OA].

the ILD error is low (typically below 1 dB), the ITD was the dominant cue (4° localization bias for a target at +25°, $B_{ITD} = 0.05$ ms and $V_{ITD} = 0.15$ ms).

It should be clear, however, that the objective criteria presented here are meant as comparison indicators to provide hints about the differences observed between situations. The localization model is rather naive and does not pretend to provide an absolute predictor of localization.

### 4.2.2 Presentation of the evaluation

A joint subjective/objective evaluation of the spatial audio rendering of the SMART-$I^2$ was conducted to evaluate its consistency with 3D graphics rendering.

Two perceptual experiments were carried out :

1. A traditional azimuth localization task where stationary subjects have to identify which of a series of visual object positioned on an arc corresponds to the position of a hear sound source.
2. A spatial exploration task which focuses on the evaluation of the parallax effect rendering in which a series of visual object are positioned on a line perpendicular to the screen and subjects are instructed to wander about the experimental area to identify which of them is the source of the sound.

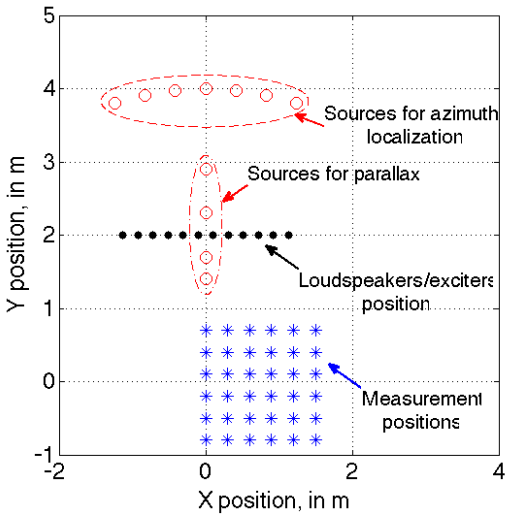



FIG. 4.7 – Top view of the installation, position of screen, audio target sources for subjective experiments and measurement points spanning the test experimental area.

In parallel to these experiments, individual exciters of the SMART-$I^2$ were measured in the experimental area using a dummy head (Neumann KU-100). Figure 4.7 presents a top view the position of the virtual sources for each experiment and the positions at which the dummy head was located. The dummy head was positioned at 36 different locations on a regular grid, grid spacing of 30 cm in both $X$ and $Y$, so as to cover the entire experimental area in which subjects could move during the experiments.

Bias and variability criteria are computed for all measured listening positions and virtual sources used in the subjective experiments. The artificial head was always facing the screen, orientated towards the $Y$ axis of figure 4.7.

These criteria are also computed for a virtual WFS setup composed of 12 ideal omnidirectional loudspeakers in an anechoic environment. This simulates an “ideal" WFS setup using loudspeakers at the same position as the exciters in the SMART-$I^2$ based on previous anechoic measurements of the dummy head. As such, our results can be compared with those of Sanson *et al.* [Cor08OA] and any observed innaccuracies could be attributed either to known WFS deficiencies (diffraction, aliasing) or to the specific radiation characteristics of the large MAP.

The subjective experiment is organised in 3 parts performed during one 40 minute session :

1. The test scene.
2. The azimuth localization task.
3. The parallax effect validation task.

Fourteen subjects (9 male and 5 female) ranging from 22 to 53 years old participated in the experiment. No subject was visually or auditory impaired. For half of the subjects, the parallax effect validation task was performed before the azimuth localization task in order to evaluate the influence of learning or habituation on the results of either task. Analysis confirmed that there was no significant effect on the results due to task order. The introductory scene contained three virtual visual loudspeakers placed

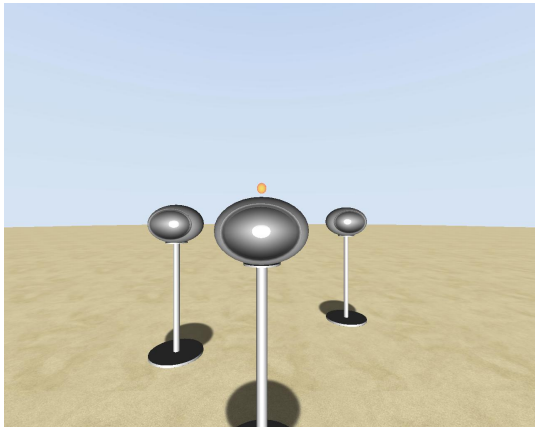


FIG. 4.8 – Introductory scene.

on a neutral background (see figure 4.8). Only one, indicated by the orange cursor above, is activate and plays music. The subject can choose which loudspeaker is active using a Wiimote. The subject is also free to move within the rendering area to get accustomed to tracked stereoscopy and WFS. During this introductory scene, the stereoscopic rendering parameters were adjusted to the individual subject.

## 4.3 Azimuth Localization Evaluation

### 4.3.1 Experimental Protocol

The goal of this experiment is to evaluate the localization accuracy of the spatial audio rendering of the SMART-$I^2$. The audio-visual scene contains 17 virtual visual loudspeakers distributed on an arc and at 4 m from the audio-visual origin (see figure 4.9). Virtual loudspeakers are separated by 3° which provides a 48° angular aperture.

The position of the visual objects has been chosen to be smaller or equal to the human audio azimuth resolution (i.e. 3-4°, see [Bla97]). Moreover, all visual objects are similar. This creates multiple visual distractors that limit bias due to audio-visual integration, the so-called ventriloquism effect [Ber98].

A single white noise burst of 150 ms with 10 ms onset and offset (Hanning window) was presented at nominal level of 68 dBA from one of the 7 target sources, indicated in figure 4.9 (background noise level at subject position was 40 dBA). The subject's task was to indicate from which of the 17 visual objects the sound has been emitted. The subject had no prior knowledge of possible target locations.

At the beginning of the experiment, the subject was positioned at the starting point (see 4.2), and instructed to remain there. There were 15 repetitions of the 7 potential positions corresponding to a total number of 105 trials. At the beginning of each trial, a position is randomly picked. Each trial is organized in the following way :

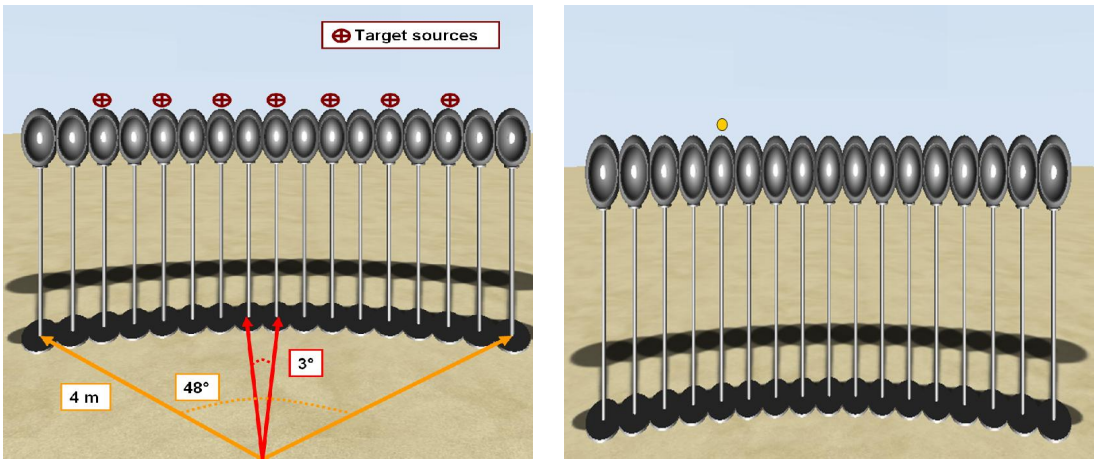



FIG. 4.9 – The audio-visual scene for azimuth localization accuracy. On the left, the 7 potential acoustic target positions and dimensions are indicated here. On the right, the scene as viewed by the user (step 3).

1. The subject indicates that he is ready by pressing a button.
2. His position and head orientation are checked with the tracking engine.
3. If the subject is ready and his head orientation is correct, the sound stimulus is played once.
4. A visual selector appears (cf. figure 4.9), and the subject indicates the perceived location of the sound source by moving the cursor on top of the corresponding visual object.

### 4.3.2 Results of the subjective evaluation

Each trial was performed in a mean time of 5.8±3.8 sec. Angles are given clockwise. A Lilliefors test rejected the null hypothesis that the empirical distribution (result values) follows a normal distribution. The Lilliefors test is a Kolmogorov-Smirnov type test where the empirical distribution function is compared to a normal distribution with the mean and standard deviation equal to the empirical ones. Therefore, it was decided to use a quantile repartition for the analysis of the results which is presented in figure 4.10. The displayed boxes have lines at the lower quartile, median, and upper quartile values. As a measure of variability, we chose the half inter-quartile range ($HIQR = (P(75\%) - P(25\%))/2$) where $P(X)$ is the percentile function which gives the value in the data set under which are found $X\%$ of the values. The half inter-quartile range is presented since its interpretation is close to the standard deviation used in literature in order to facilitate comparisons with previous localization studies. The

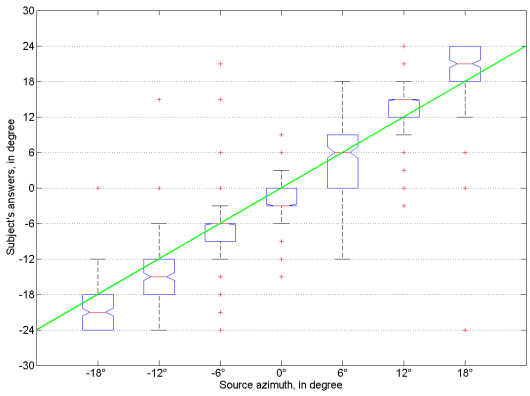



FIG. 4.10 – Subjects answers for the target sources.

whiskers displayed in figure 4.10 (lines extending from each end of the boxes) show the extreme values beyond the lower and the upper quartiles. Outliers are data with values beyond the ends of the whiskers and are indicated with a red '+'. The notches represent an estimate of the uncertainty about the medians. The different target sources can thus be considered to be well identified by subjects in this task since notches are non-overlapping in figure 4.10.

The median of the error is always smaller or equal to 3°, which is the angular separation between two neighbouring speakers. For extremal target sources on the left the median error is close to -3°, and for the ones on the right to 3°. This denotes a tendency to exagerate the angular position of lateral sources. The $HIQR$ is always lower than 3°, except for the source placed at 6° for which it reaches 4°.

### 4.3.3 Results of the objective evaluation

The bias and variability for ITD and ILD are displayed in figures 4.11 at the position of the subjects for all sources used in the localization test.

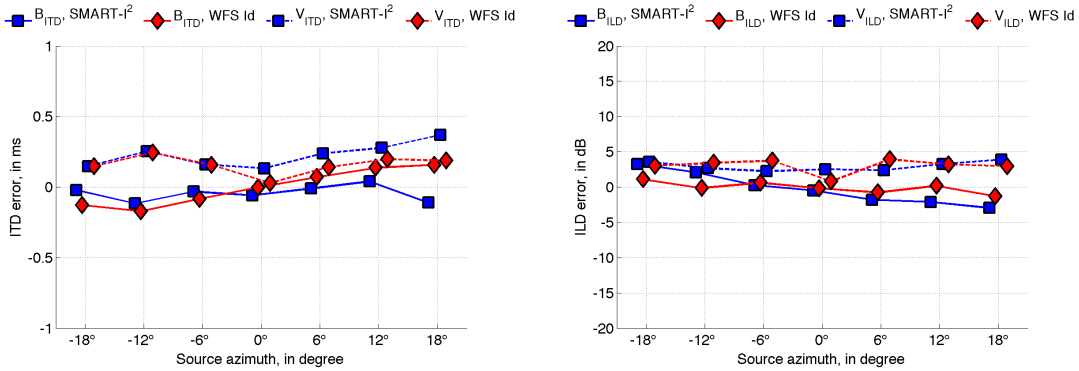



FIG. 4.11 – $B_{ITD}$ and $V_{ITD}$ (on the left), $B_{ILD}$ and $V_{ILD}$ (on the right) for SMART-$I^2$ and ideal loudspeakers ($WFSId$), center position.

The errors for localization cues remain relatively small for the SMART-$I^2$ and are comparable to errors obtained with the ideal WFS setup. However, the variability of the ITD is slightly higher for the SMART-$I^2$ than for the ideal WFS setup. The ILD bias exhibits a curve which would suggest a compression of the perceived azimuth of virtual sources. This is an opposite effect that was noticed in the subjective test indicating a moderate dilatation of the perceived azimuth.

### 4.3.4 Discussion

The azimuth localization task was reported by the users to be quite easy and intuitive. In this section, the results are discussed and compared with similar previous studies. The validity of localization in the complete listening area is verified further on.

#### Comparison with other studies

The absolute human localization blur, according to [Bla97], is between 1° and 4° in front of the listener, depending on the test signal and on the reporting method. The absolute localization bias of the present study is thus very close to human ability.

The observed extra-lateralization has no precise explanation. It should be recalled that the localization model used in the objective part of this study is rather naive. It does not account for the precedence effect. The angular localization is indeed biased in our environment. This could be due to residual reflections from the side wall which cannot be properly taken into account in the model, but this rationale does not fully explain the results. Interestingly, this effect has also been reported to the same extent (2° to 3° over-lateralization at ±20° or 30°) in headphone-based localization experiments

using individual HRTFs [Wig89][Bro95]. As such, one cannot claim this as a system default, but could be more linked to some perceptual critera.

Verheijen achieved localization experiments on a WFS setup composed of 24 loudspeakers with an 11 cm spacing using a similar method where the subjects had to indicate which of the visible (real) loudspeakers was emitting sound for both synthetised sources, using WFS, and real sources.[Ver97] Verheijen also tested a loudspeaker spacing of 22 cm using thus only 12 loudspeakers as in the current study. Verheijen reported mean localization error for synthesized sources of approximately 3.5° independent of loudspeaker spacing. The associated standard deviation was 1.5°. Localization accuracy for real sources was very similar.

The localization accuracy provided by the SMART-$I^2$ is comparable to these results for a similar WFS configuration. However, the $HIQR$ obtained with the SMART-$I^2$ is higher than the standard deviation shown there. This may be attributed to the radiation characteristics of the MAP loudspeakers as compared to those of electrodynamic loudspeakers as used by Verheijen. It is expected that the use of advanced multichannel equalization techniques [Cor06ME] will increase the localization accuracy in SMART-$I^2$. This will be a topic for further studies.

The SMART-$I^2$ results should also be compared to results provided by other sound rendering technologies. For an ideal fourth order ambisonic rendering, with a circular 12 loudspeaker array, Bertet *et al.* reported a median error is between 1° and 3° and a $HIQR$ between 2° and 5° at the sweet spot.[Ber07] According to [Ngu08], a binaural rendering using individual HRTFs showed an average absolute localization error of 0° to 3.5° depending on source azimuth (from $-30°$ to $+30°$ every 10°) with a standard deviation of 6° to 9°. The latter study also included a combined audio-visual study which showed that audio-visual integration occurs and that there is a dominance of visual localization over auditory localization which removes the bias and reduced the standard deviation to 2.1° in bimodal conditions. A similar effect can be expected in the SMART-$I^2$ system for congruent audio-visual objects.

#### Extension to entire listening area

Figures 4.12 present ITD and ILD bias (solid line) and variability (dashed line) criteria averaged over all measured positions (cf. figure 4.7). Whiskers indicate standard deviation.

Averaged ITD and ILD bias and variability criteria follow similar tendencies as those observed for the center position. Moreover, the standard deviation of these criteria remains low, typically below 0.1 ms for ITD based criteria and below 1 dB for ILD based criteria. It can thus be expected that localization remains almost unbiased over the entire listening area with a reduced localization accuracy for side virtual sources as in the center listening position. Localization accuracy can thus be expected to be at least as good as sweet spot based techniques such as fourth order ambisonics but over an extended listening area.

## 4.4 Parallax effect evaluation

### 4.4.1 Experimental protocol

The audio-visual scene contains 8 “visual" loudspeakers arranged in a line perpendicular to the screen at 30 cm intervals. The first loudspeaker is 1.1 m away from the audio-visual origin (see figure 4.13), resulting in the three first loudspeakers being in front of the screen. Only 4 of the visual target sources are used as audio targets in this session (1.4 m, 1.7 m, 2.3 m and 2.9 m). The audio stimulus is a 1 sec white noise burst, low pass filtered at 4 kHz, on which a 18 Hz amplitude modulation is applied (70 dBA presentation level at the audio-visual origin) followed by a 1 sec silence. The stimulus was played continuously until the subjects made their selection. The sound levels of the different sources were chosen in order to have the same effective level at the audio-visual origin. Parallax was thus the only available cue to the subjects for distance estimation. There were 10 repetitions of each of the 4 target positions corresponding to a total number of 40 trials. In this experiment, subjects were instructed to move about within the defined experimental area (see figure 4.2). Each trial begins with

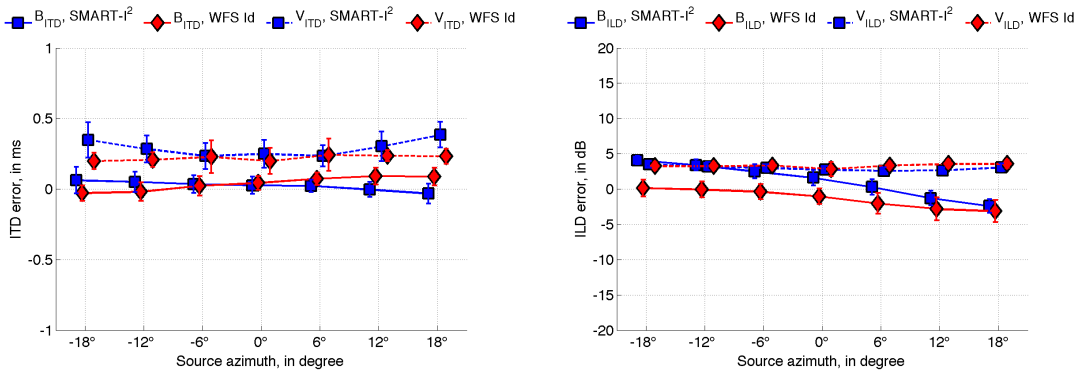



FIG. 4.12 – $B_{ITD}$ and $V_{ITD}$ (on the left), $B_{ILD}$ and $V_{ILD}$ (on the right) for SMART-$I^2$ and ideal loudspeakers ($WFSId$), average over all measured positions.

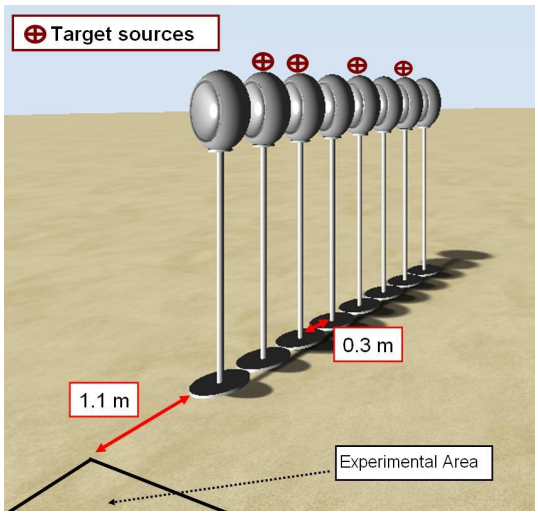



FIG. 4.13 – The audio-visual scene for parallax effect evaluation in the SMART-$I^2$.

the user located at the starting point. Subjects are instructed to move slowly from this point to the right of the listening area. The experimental procedure was the same as previously described. Subject position was checked before the beginning of the trial but not the head orientation.

### 4.4.2 Results of the subjective evaluation

The results for parallax evaluation are presented in figures 4.14 and 4.14 in the same manner than the previous results. The average response time for each trial is $23 \pm 13$ sec with little variation between source positions. In figure 4.14, the different data sets corresponding to the target sources are not clearly separated. The second source (1.7 m : focused source) is perceived to be at the relative position of the third source (2.3 m : source behind the screen). The two other sources are well identified. The $HIQR$ in meters can be seen in figure 4.14. It is minimum for the closest source (1.4 m) where it reaches 0.15 m. The $HIQR$ is 0.45 m for the source at 1.7 m, and approximately 0.3 m for the two others.

Figure 4.14 presents the median error and the $HIQR$ converted into degrees as perceived at the extreme right listening position ($x = 1.5$ m, $y = 0$ m) which is the end of the path proposed in the instructions. This is the position in the experimental area (cf. figure 4.2) at which there is the maximum angular difference between visual objects. Therefore, this is the position where the subjects

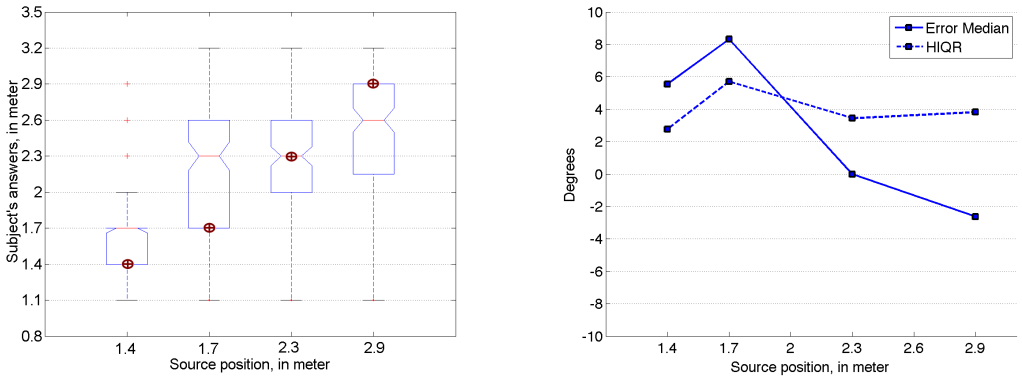



FIG. 4.14 – On the left, subject responses for the target source distance position and target locations. On the right, median error and $HIQR$ for the target sources. Errors are indicated in degrees as perceived at the extreme right listening position ($x = 1.5$ m, $y = 0$ m).

could most easily perform the required task.

Figure 4.14 shows that the median error is positive for the two first focused sources, null for the third one, and negative for the last one. This indicated that the two closest sources are biased to the right (further distance), the third source (at 2.3 m) is accurately localized whereas the last one is perceived too much to the left (closer to the screen location). The $HIQR$ is close to 4° for all sources which is similar to the results of the first experiment. The estimation of degrees at this position provides an explaination for the large variation of $HIQR$ observed in figure 4.14.

### 4.4.3 Results of the objective evaluation

During the experiment, subjects were instructed to follow the path between the center listening and an extreme side listening position ($x = 1.5$ m, $y = 0$ m in figure 4.7). At the center listening position, the ITD and ILD bias and variability are almost independent on the source distance and are therefore not presented here. At this position, ITD and ILD bias are in the range of $0 \pm 0.05$ ms for both ideal WFS and SMART-$I^2$, $0 \pm 0.5$ dB for ideal WFS and $-1.5 \pm 0.5$ dB for SMART-$I^2$. The small observed ILD bias for SMART-$I^2$ might be due to the presence of the wall in the current installation.

Objective criterion at the extreme right position ($x = 1.5$ m, $y = -0.5$ m in figure 4.7) are presented in figures 4.15. It can be seen that there is a positive bias ($B_{ITD} \simeq 0.2$ ms and $B_{ILD} \simeq 7$ dB) for sources located in front of the screen (distance lower than 2 m) synthesized using SMART-$I^2$. This bias is not observed for ideal WFS. The consistent bias for both ITD and ILD might explain the overestimation of perceived distance in the subjective experiment. This might cause a localization bias to the right of the perceived azimuth of the source compared to the target azimuth.

For sources located behind the array, a positive bias is observed for $ILD$ ($B_{ILD} \simeq 5$ dB) whereas $B_{ITD}$ is close to null. This may explain the limited bias in localization accuracy observed for these sources accounting for the dominance of ITD for large band stimuli [Wig92].

### 4.4.4 Discussion

The subjects reported this task very difficult to perform. The subjects were all a bit disturbed by the missing “static" acoustical cues for distance perception (level differences, room effect, ...). During the first trials, subjects strategies were to explore the entire rendering area trying to find the best localization place. After that, 80% of the subjects went directly to the extreme right listening position to choose the loudspeaker, and the other part still proceed to a systematic exploration or to a random walking among the rendering area.

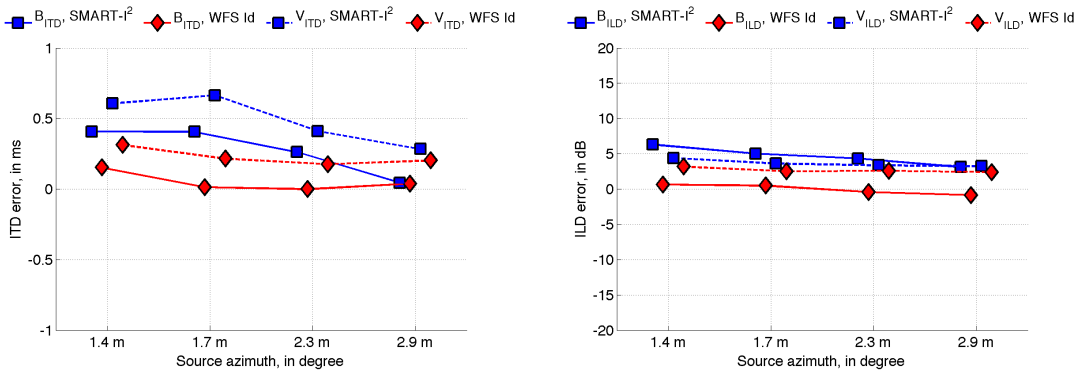



FIG. 4.15 – $B_{ITD}$ and $V_{ITD}$ (on the left), $B_{ILD}$ and $V_{ILD}$ (on the right) for SMART-$I^2$ and ideal loudspeakers ($WFSId$), extreme right listening position.

The motion parralax is a dynamic distance cue which is known to be not as salient as static distance cues.[Spe93][Ash95] In [Spe93], blind-folded subjects were instructed to locate a sound source in a very similar way than our experiment. A number of loudspeakers were located on a line passing through the starting position at 2 m to 6 m from the starting position. The subjects were asked to walk either 2 m or 4 m in a direction perpendicular to this line while one loudspeakers was active (synthesized voice). The subjects were then instructed to report the exact location of the active source by walking to it. The results indicate an over-estimation of the distance of the closer sources and an under-estimation of the distant ones. The same phenomenom has been noticed in [Ash95]. The standard deviation found in [Spe93] was approximately 0.6 m, greater than the $HIQR$ of our experiment. However, the reporting method we used is quite different than in [Spe93] and [Ash95]. The current study task can almost be considered as a localization task with visual cues seen from multiple listening positions.

The objective analysis revealed inconsistent localization cues which explain the localization bias for sources in front of the screen. In further studies, these errors should be compensated for using more advanced equalization techniques or by modifying the acoustical characteristics of the panel. Considering the furthest position (2.9 m), the compression in distance might also be due to the “natural" room effect of the real environment. The latter provides distance cues related to direct over reverberant energy which may bias distance perception towards the location of the screen.

From a visual point of view, a compression of distance perception has been noticed for a while in immersive virtual environments. It does not seem to be linked to the field of view and binocular viewing, and for the moment no convincing explanations has been found to this phenomenom.[Cre05] In the SMART-$I^2$, this can contribute to the mislocalization of the audio-visual objects in terms of distance.

In conclusion, it can be stated that the SMART-$I^2$'s ability to render the audio-visual motion parralax effect is similar to the one observed with real sources. Therefore, it can be expected that the results would be improved using consistent level differences and virtual room effect. Audio-visual integration should also improve the precision of the motion parallax effect rendering in a similar way than for the localization accuracy for static users (cf discussion in section 4.3.4).

## General discussion and conclusion

In this paper, an immersive audio-visual environment, the SMART-$I^2$, has been presented and evaluated. The design focus was placed upon audio-visual consistency which may contribute to an increased feeling of presence in the virtual environment.

Subjects were globally satisfied with the proposed rendering. Further studies should be made with this system with more realistic audio-visual scenes to evaluate more precisely the perceptual realism

provided by the SMART-$I^2$. A more complete acoustical study of the large MAP is also required in order to increase the precision of the WFS rendering, especially for sources located in front of the screen.

The immersion quality was reported by the subjects to be sufficient and that they really felt they were “in the scene". One fact that has been reported was that the limited size of the audio-visual window caused some disturbances in extremal positions. Installation of the second MAP will offer users a larger field of view and thus should raise up the immersion quality.

In these experiments, the level of interaction was very limited in terms of direct interaction with the scene and also interactions between multiple users. The addition of both may contribute to an increase in the sense of presence. This will be verified in further studies.

The global feeling about the audio-visual rendering provided by the SMART-$I^2$ was that the consistency between audio and visual features is very accurate and that immersion quality is already convincing. The SMART-$I^2$ was perceived as an audio-visual window opening into a plausible and consistent virtual world. The results of the objective analysis and subjective evaluation confirm that point. The localisation accuracy shown by subjects permits a global validation of the WFS rendering, the visual rendering and the audio-visual calibration.

# Chapitre 5

# Conclusion

## 5.1 Bilan du stage

### 5.1.1 Travail réalisé

Ce stage aura donc été l'occasion, par la création du SMART-$I^2$, de m'intéresser à des thèmes divers. Dans un premier temps, une étude acoustique cherchant à étudier l'influence de l'encastrement sur le rayonnement d'un MAP et à quantifier la part de non-linéarités présente dans ce rayonnement a été menée. Ensuite, la réalisation à proprement parler du dispositif, ainsi que sa calibration pour que le rendu proposé soit cohérent avec le monde physique ont été réalisés. Enfin, le rendu audio-visuel proposé par le SMART-$I^2$ a été validé en utilisant de façon croisée une étude acoustique objective et une étude audio-visuelle perceptive.

### 5.1.2 Bilan personnel

Au cours de ce stage, j'ai eu l'occasion d'en apprendre beaucoup dans des domaines très variés allant de l'informatique graphique au maniement de la perceuse sur de l'aluminium en passant l'identification de modèles non-linéaires. Ce stage m'a aussi permis d'avoir un contact plus poussé avec le monde de la recherche. En effet, au cours des 5 mois et demi passés au LIMSI, j'ai pu me consacrer entièrement à cette activité, faire des choix concrets et voir ou me menaient ces choix par la suite. De plus, la co-rédaction d'un article en anglais a été une activité elle aussi très formatrice sur le plan de la recherche.

## 5.2 Perspectives de recherche

La réalisation de ce prototype a permis de soulever de nombreuses questions qui ouvrent maintenant de nouvelles perspectives de recherche.

### 5.2.1 A court terme

#### Généralisation de l'outil d'étude des non-linéarités

Une validation de la méthode développée pour analyser les non-linéarités présentes dans le rayonnement va être réalisée. Cet outil, s'il s'avère performant, pourra alors être interéssant pour l'étude fine de différents haut-parleurs dans le but de les comparer.

#### Achèvement du protoype

Pour l'instant, le prototype n'est pas encore achevé. Le montage du second écran est donc prévu prochainement ce qui permettra d'élargir les champs audio et visuel utiles. Nous serons alors à même de créer des scènes audio-visuelles plus complètes. De plus, l'implémentation d'un effet de salle et la prise en compte de l'atténuation du son lors de sa propagation dans l'air sont aussi à terminer.

#### Aspects "interaction" et "multi-utilisateurs"

L'aspect interactif du prototype n'a pas encore été utilisé ni évalué. Une scène interactive, de type "flipper 3D" va donc être développée pour permettre d'étudier la façon dont les utilisateurs interagissent avec le dispositif. L'aspect multi-utilisateur n'a lui non plus pas été validé. Nous comptons donc développer par la même occasion des stratégies permettant à deux utilisateurs de partager l'espace audio-visuel avec le SMART-$I^2$ et étudier la façon dont les utilisateurs communiquent lors de leur immersion.

### 5.2.2 A plus long terme

#### Optimisation acoustique du dispositif

Une caractérisation acoustique précise de la structure complète sera effectuée pour calculer des filtres d'inversion qui soient les plus performant possibles et pour détecter les éventuels défauts et limites du système d'un point de vue acoustique. Puis l'amélioration du rayonnement acoustique individuel de chacun des excitateurs sera envisagé, par l'utilisation de techniques de traitement du signal adaptées et le placement de matériaux absorbants. Une modélisation du panneau dans le but de son optimisation est aussi envisagée.

#### Interactions entre l'ouïe et la vue

Des expérimentations cognitives concernant la cohérence audio-visuelle et l'influence réciproque qu'ont l'ouïe et la vue l'une sur l'autre, dans ce contexte, seront ensuite réalisées. Le but de ces expériences serait d'évaluer le degré de précision audio et visuel qu'il est nécéssaire d'atteindre pour que le niveau de réalisme perceptuel soit suffisant et donc que le rendu soit convaincant. Ainsi, si une marge d'erreur humainement non-perçue au niveau audio-visuel est mise en évidence, des simplifications sur les algorithmes de calcul pourront être réalisées.

# Bibliographie


[Ash95] Ashmead D.H., Deford L.D., Northington A. "Contribution of listener's approaching motion to auditory distance perception", *Journal of Experimental Psychology : Human Perception and Performance*, **21**(2),pp. 239-256, 1995.

[Beg94] Begault D., "3D sound for virtual reality and multimedia", Cambridge, MA : Academic Press, 1994.

[Ber93] Berkhout A. J. , de Vries D. and Vogel P., "Acoustic Control By Wave Field Synthesis", *Journal of Acoustical Society of America*, vol. 93, pp 2764-2778, 1993

[Ber98] Bertelson P., Aschersleben G., "Automatic visual bias of perceived auditory location", *Psychonomic Bulletin & Review*, **5**(3), pp. 482-489, 1998.

[Ber07] Bertet S., Daniel J., Parizet E., Gros L., Warusfel O., "Investigation of the perceived spatial resolution of higher ambisonics sound fields : a subjective evaluation involving virtual and real 3D microphones", *30th International Conference of the Audio Engineering Society*, Saarileskä, Finland, 2007.

[Bla97] Blauert J., "Spatial Hearing, the Psychophysics of Human Sound Localization", MIT Press, first published in 1974, re-edited in 1997.

[Boo04] Boone M. M., "Multi-actuator Panels as loudspeakers arrays for Wave Field Synthesis", *Journal of the Audio Engineering Society*, **52**(7/8), pp. 712-723, July-August 2004.

[Bou08] Bouchara T., "Le Scene-Modeler : Des outils pour la modélisation de contenus multimedia spatialisés." *Actes des 13 èmes Journées d'Informatique Musicale*, Albi, France, 2008. http ://gmea.net/upload/

[Bro95] Bronkhorst A., "Localization of real and virtual sound sources", *The Journal of the Acoustical Society of America*, **98**(1), pp. 2542-2553, 1995.

[Bru04] de Bruijn W., "Application of Wave Field Synthesis for life-size videoconferencing" *Phd thesis*, Delft University of Technology, 2004.

[Bur87] Burbeck S., "Applications Programming in Smalltalk-80(TM) : How to use Model-View-Controller (MVC)" *Ph.D. thesis*, University of Illinois at Urbana-Champaign, 1987.

[Cho06] Chou D., "Efficacy of Hammerstein Models in Capturing the Dynamics of Isometric Muscle Stimulated at Various Frequencies", *Ph.D. Thesis*, Massachussetts Institute of Technology, USA, 2006.

[Cor04] Corteel E. *Caractérisation et Extensions de la Wave Field Synthesis en conditions réelles* PhD Thesis, Université Paris 6, 2004.

[Cor06ME] Corteel E., "Equalization in an extended area using multichannel inversion and wave field synthesis", *Journal of the Audio Engineering Society*, **54**(12), pp. 1140-1161, December 2006

[Cor06AL] Corteel E. "On the use of irregularly spaced loudspeaker arrays for Wave Field Synthesis, potential impact on spatial aliasing frequency" *9th Int. Conference on Digital Audio Effects (DAFx-06)*, Montreal, Canada, 2006.

[Cor07] Corteel E., "Objective and subjective comparison of electrodynamic and MAP loudspeakers for Wave Field Synthesis" *30th International Conference of the Audio Engineering Society*, Saariselkä, Finland, 2007.

[Cor08SW] Corteel E., Pellegrini R., Kuhn-Rahloff C., "Wave Field Synthesis with Increased Aliasing Frequency", *124th Convention of the Audio Engineering Society*, Amsterdam, Netherland, 2008.

[Cor08OA] Sanson J., Corteel E., Warusfel O., "Objective and subjective analysis of localization accuracy in Wave Field Synthesis", *124th Convention of the Audio Engineering Society*, Amsterdam, Netherland, 2008.

[Cre05] Creem-Regehr S. H., Willemsen P., Gooch A. A., Thompson W. B. "The influence of restricted viewing conditions on egocentric distance perception : Implications for real and virtual indoor environments", *Perception* **34**(2), pp. 191-204, 2005.

[Cru92] Cruz-Neira C., Sandin D., DeFanti T., Kenyon R., Hart J., "The CAVE-Audio visual experience automatic virtual environment", *Communications of ACM*, **35**(6), 64-72, 1992

[Dan00] Daniel J., "Représentation de champs acoustiques, application à la reproduction et à la transmission de scènes sonores complexes dans un contexte multimédia", *Ph.D. thesis*, University of Paris 6, Paris, France, 2000.

[EVE] EVE project homepage (Experimental Virtual Environement) http ://eve.hut.fi/

[Far00] Farina A. "Simultaneous measurement of impulse response and distortion with a swept-sine technique", *108th Convention of the Audio Engineering Society*, Paris, France, 2000.

[Fuc01] Fuchs P. , Moreau G., Papin J.P. *Traité de la réalité virtuelle* Les Presses de l'Ecole des Mines, 2001.

[Fro05] Fröhlich B., Blach R., Stefani O., "Implementing Multi-Viewer Stereo Displays", *WSCG Conference Proceedings*, Plzen, Czech Republic, 2005.

[Ger73] Gerzon M. A., "Periphony : With-Height Sound Reproduction", *Journal of the Audio Engineering Society*, **21**(1), pp. 2-10, 1973.

[Ham30] Hammerstein A., "Nichtlineare integralgleichung nebst anwendungen", *Acta Mathematica*, 54, pp. 117-176, 1930.

[Has99] Hasler M., "Phénomènes non linéaires, Chapitre 3 : Séries de Volterra", *EPFL Lausanne*, 1999.

[Jac04] Jacquemin C., "Architecture and Experiments in networked 3D Audio/Graphic rendering with Virtual Choreographer", *Proceedings of Sounds and Music Computing*, Paris, France, 2004.

[Kus06] Kuster M., De Vries D., Beer D., Brix S., "Structural and Acoustic Analysis of Multiactuator Panels", *Journal of the Audio Engineering Society*, **54**(11), 2006.

[Las05] Lashkari K., "A Modified Volterra-Wiener-Hammerstein Model for Loudspeaker Precompensation", *Conference Record of the Thirty-Ninth Asilomar Conference on Signals, Systems and Computers*, 2005.

[Lom97] Lombard M. and Ditton T., "At the heart of it all : The concept of presence", *Journal of Computer-Mediated Communication*, **3**(2), September 1997.

[Lok07] Gröhn M., Lokki T., and Takala T., "Localizing sound sources in a cave-like virtual environment with loudspeaker array reproduction", *Presence : Teleoperators & Virtual Environments*, **16**(2), pp. 157-171, April 2007.

[MAX] Zicarelli D., Taylor G., Clayton J.K., Dudas R., Nevil B., "MAX 4.6 : Reference Manual", http ://www.cylcing74.com

[Ngu08] Nguyen K. V., Suied C., Viaud-Delmon I., Warusfel O., "Intergrating visual and auditory spatial cues in a virtual reality environment", *Submitted.*

[Pue08] Pueo B., Escolano J., Javier Lopez J., Ramos G., "On Large Multiactuator Panels for Wave Field Synthesis Applications", *124th Convention of the Audio Engineering Society*, Amsterdam, The Netherlands, 2008.

[Pul97] Pulkki V., "Virtual sound source positioning using vector base amplitude panning", *Journal of the Audio Engineering Society*, **45**(6), pp. 456-466, June 1997.

[Rug81] Rugh W. J., "Nonlinear System Theory : The Volterra/Wiener Approach" *The Johns Hopkins University Press*, 1981. URL : http ://www.ece.jhu.edu/rugh/volterra/book.pdf

[Spe93] Speigle, J. M., Loomis, J. M., "Auditory distance perception by translating observers", *Proceedings of IEEE Symposium on Research Frontiers in Virtual Reality*, San Jose, CA, October 25-26, 1993.

[Spr06] Springer J.P., Sladeczek C., Scheffler M., Hochstrate J., Melchior F., Frohlich B., "Combining Wave Field Synthesis and Multi-Viewer Stereo Displays", *Proceedings of the IEEE conference on Virtual Reality*, Washington, IEEE Computer Society, pp. 237-240, 2006.

[Ste88] Stern R. M. , Zeitberg A. S., and Trahiotis C., "Lateralization of complex binaural stimuli, a weighted-image model", *Journal of the Acoustical Society of America*, 84(1), pp. 156-165, 1988.

[Tou08] Touzé C., Camier C., Favraud G., Thomas O., "Effect of imperfections and damping on the type of non-linearity of circular plates and shallow spherical shells", *Mathematical problems in Engineering*, **2008**, 2008.

[Tur03] Turunen J., Tanttu J. T., Loula P., "Hammerstein Model for Speech Coding", *EURASIP Journal on Applied Signal Processing*, **2003**(12), pp. 1238-1249, 2003.

[Ver97] Verheijen E., "Sound reproduction by Wave Field Synthesis", *Ph.D. thesis*, Delft University of Technology, 1997.

[VIR] Jacquemin C., "Virtual Choreographer Reference Guide (version 1.4)", LIMSI-CNRS and Université Paris 11 http ://virchor.sourceforge.net

[Wes95] Westwick D. T., "Methods for the Identification of Multiple-Input Nnonlinear Systems", *Ph.D. Thesis*, Mac Gill University, Canada, 1995.

[Wig89] Wightman F. L., Kistler D. J., "Headphone simulation of free-field listening II : Psychophysical validation.", *The Journal of Acoustical Society of America*, **85**(2), pp. 868-878, 1989.

[Wig92] Wightman F. L., Kistler, D. J. "The dominant role of low-frequency interaural time differences in sound localization" *Journal of the Acoustical Society of America*, **91**(3), pp. 1648-1641, 1992.

# Annexe A

# Figures pour chacuns des systèmes étudiés

## A.1 Haut-parleur de référence

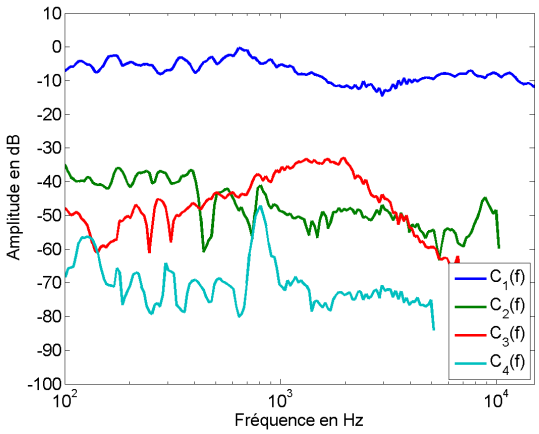



FIG. A.1 – Réponses fréquentielles dans l'axe.

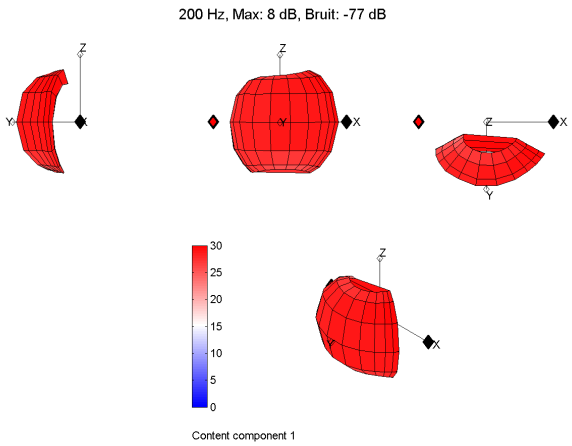



FIG. A.2 – Ecarts au modèle monopolaire en représentation 1 pour $f = 200$ Hz.

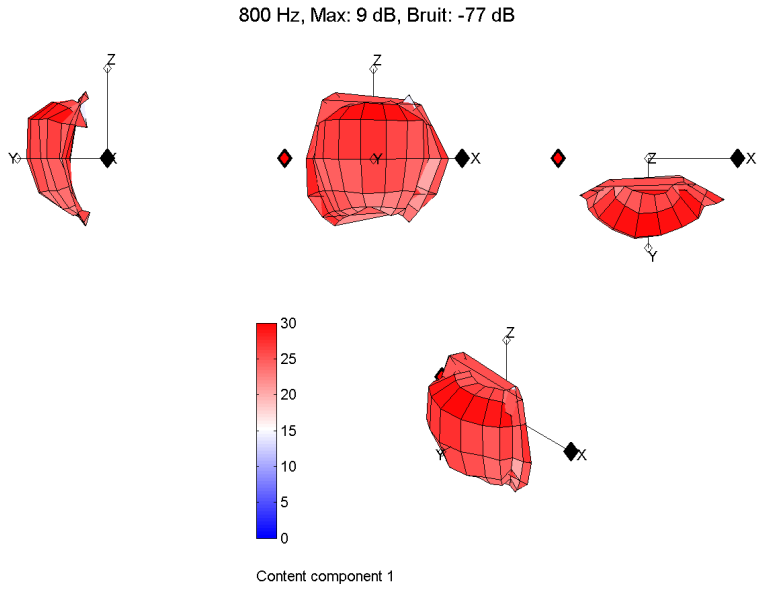



FIG. A.3 – Ecarts au modèle monopolaire en représentation 1 pour $f = 800$ Hz.

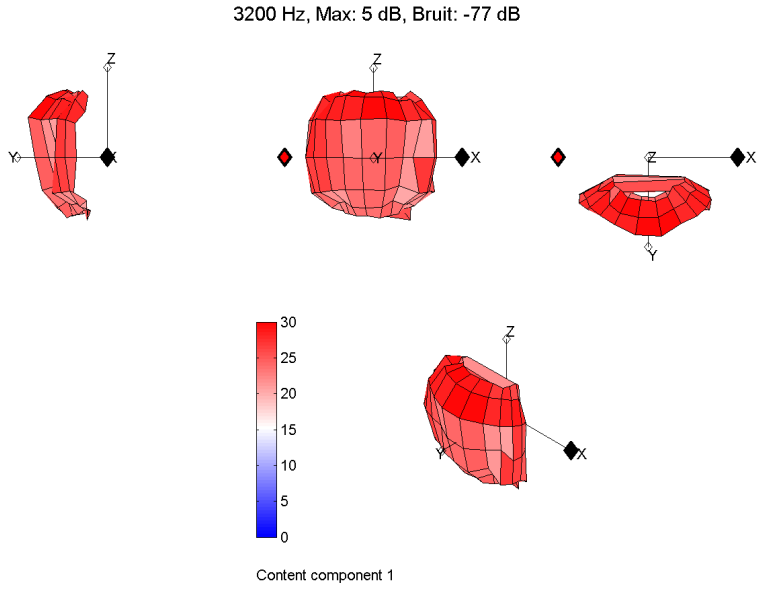



FIG. A.4 – Ecarts au modèle monopolaire en représentation 1 pour $f = 3200$ Hz.

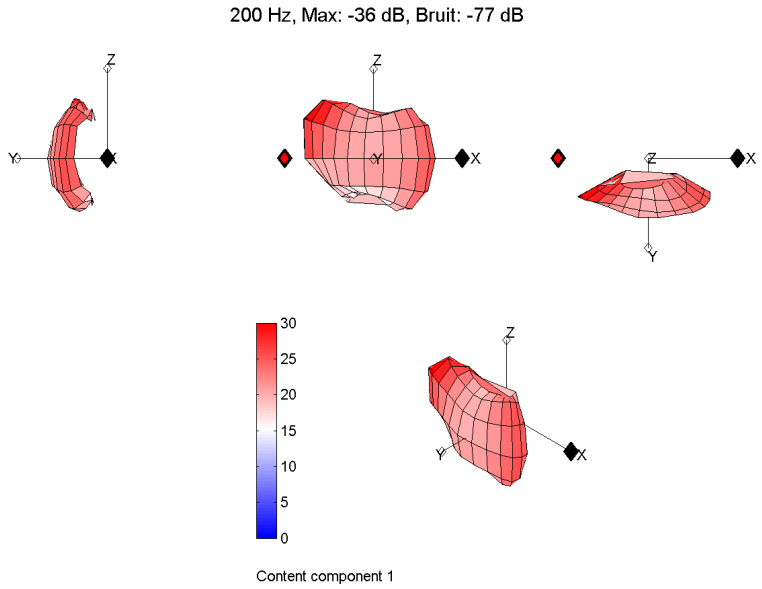



FIG. A.5 – Répartition spatiale du TDH pour $f = 200$ Hz.

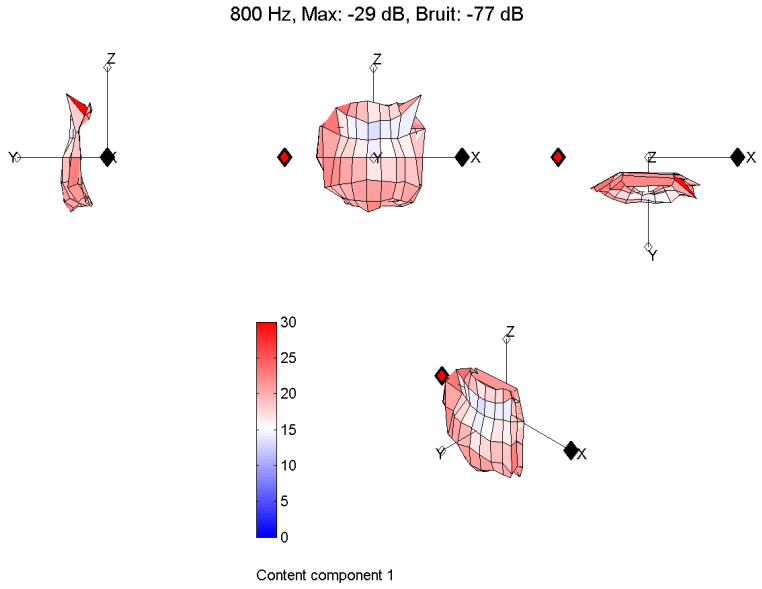



FIG. A.6 – Répartition spatiale du TDH pour $f = 800$ Hz.

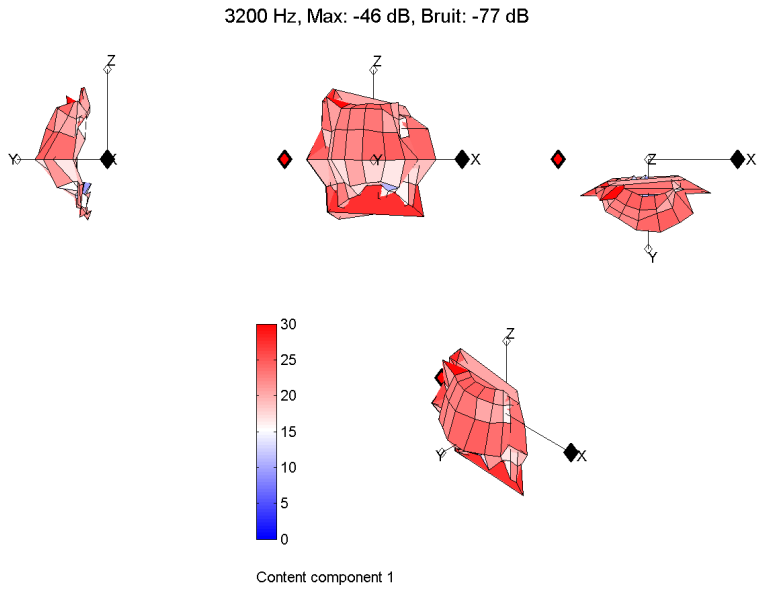



FIG. A.7 – Répartition spatiale du TDH pour $f = 3200$ Hz.

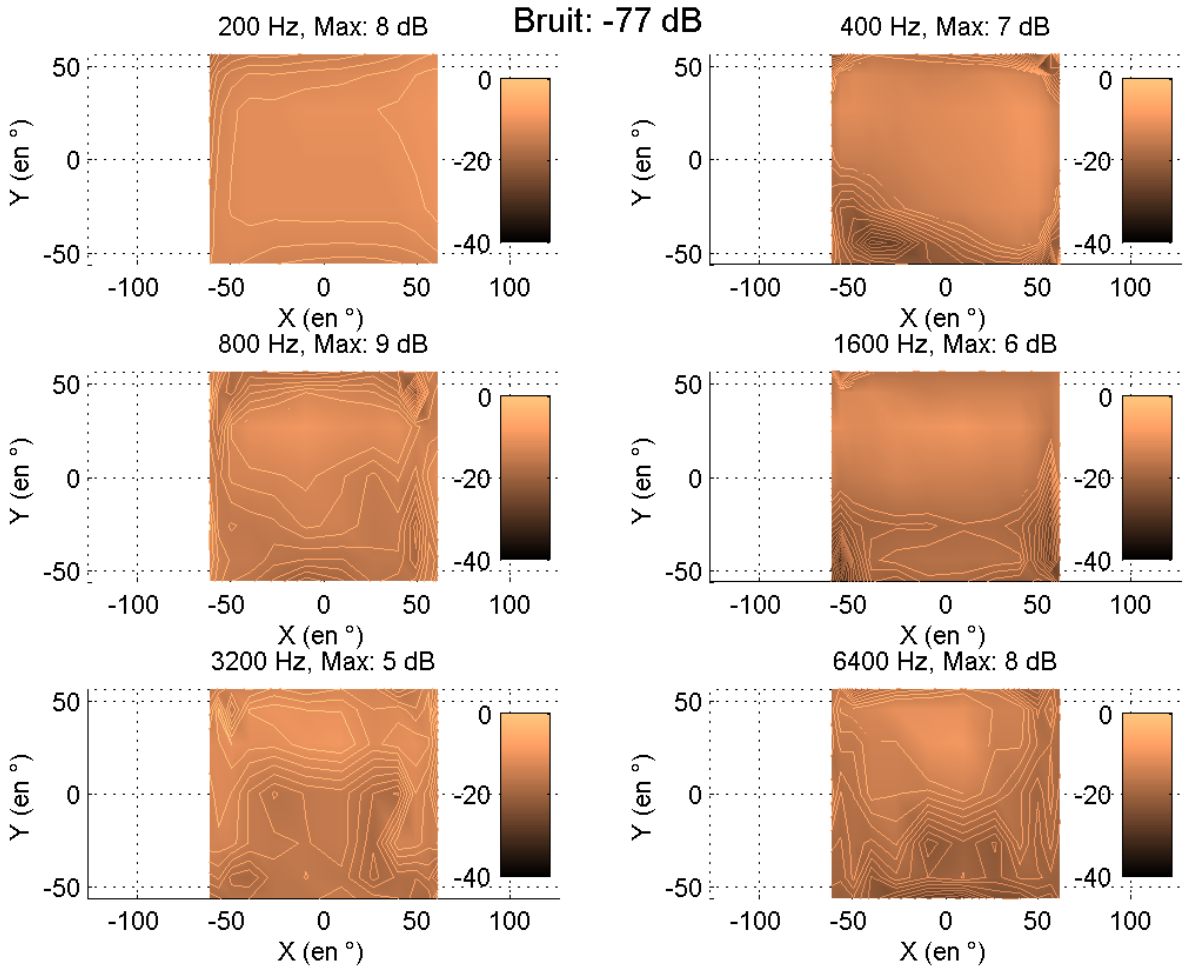



FIG. A.8 – Ecarts au modèle monopolaire en représentation 2.

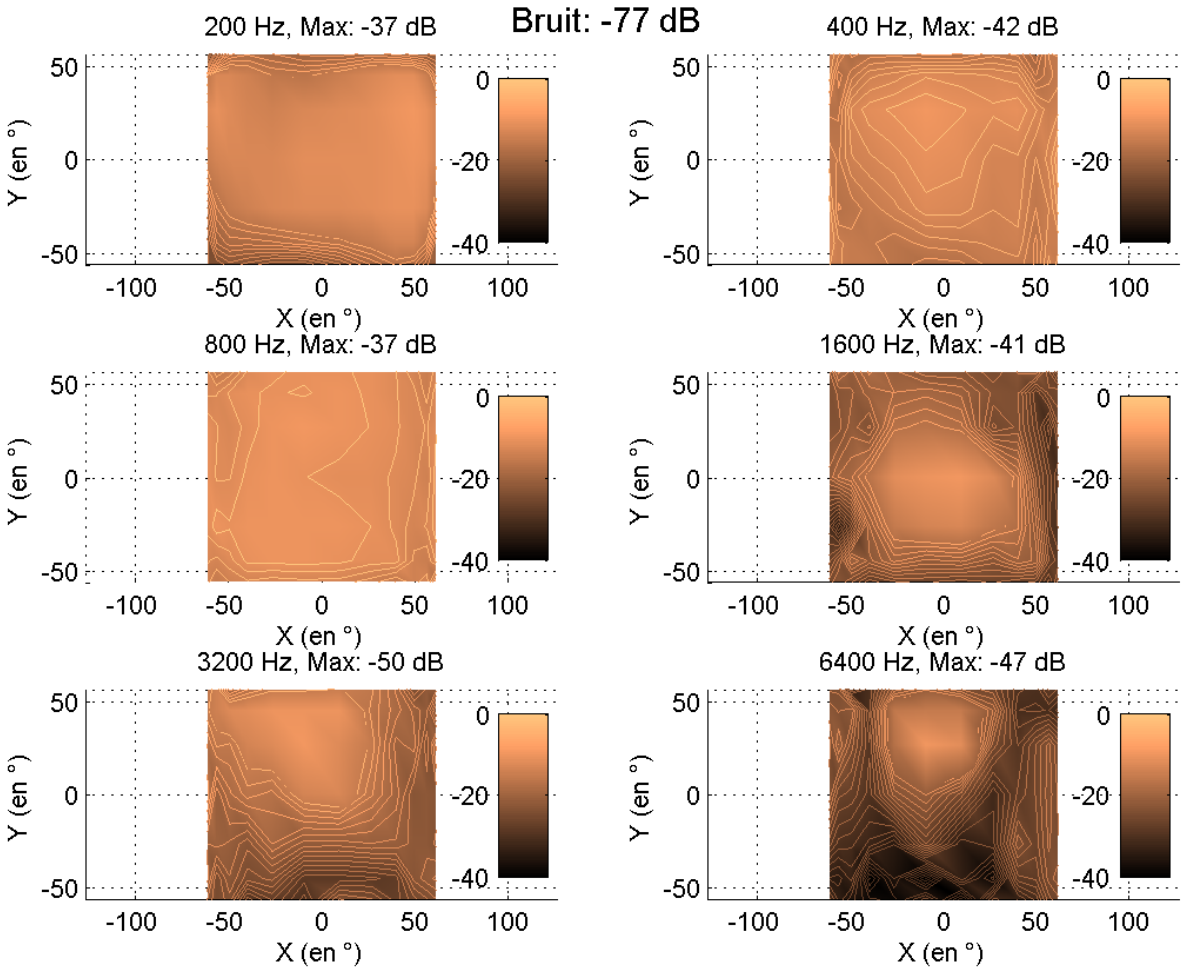



FIG. A.9 – Taux de distorsion harmonique en représentation 2.

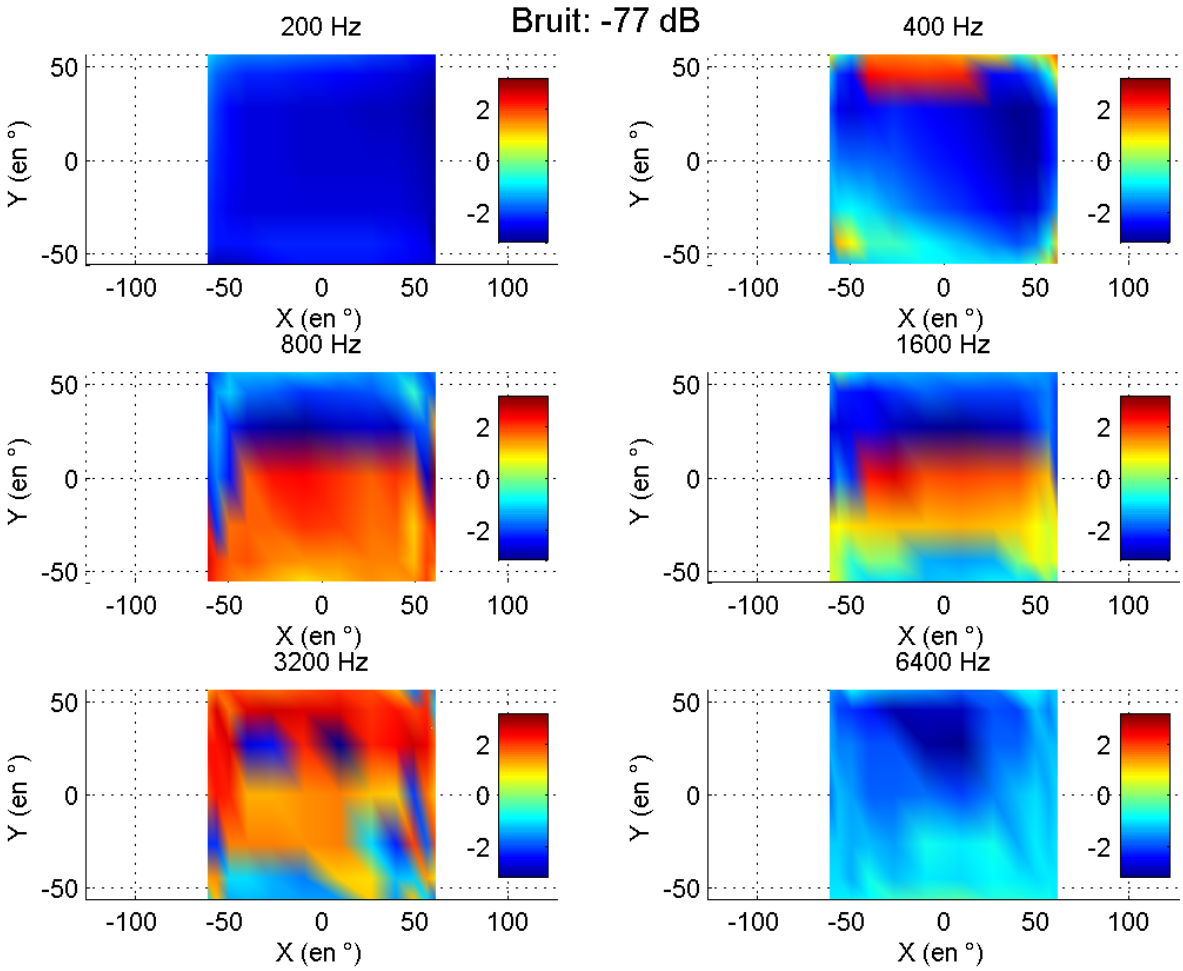



FIG. A.10 – Ecarts au modèle monopolaire en représentation 3.

## A.2 MAP libre

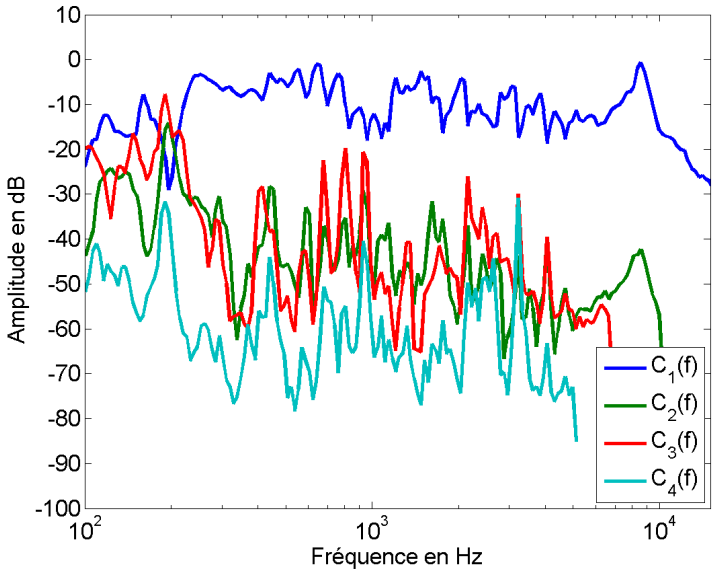



FIG. A.11 – Réponses fréquentielles dans l'axe.

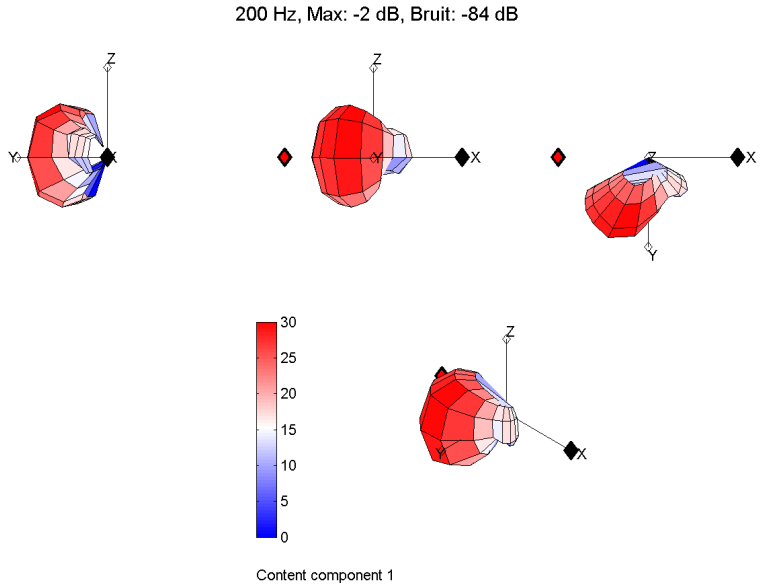



FIG. A.12 – Ecarts au modèle monopolaire en représentation 1 pour $f = 200$ Hz.

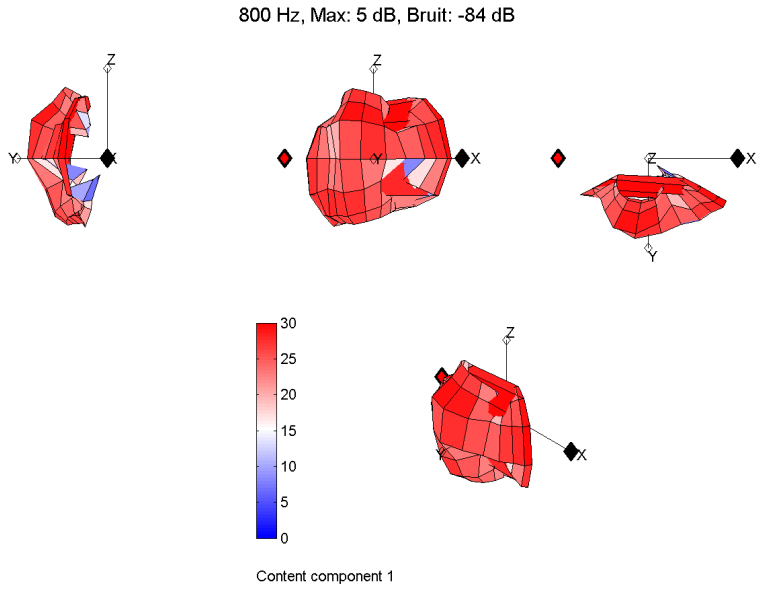



FIG. A.13 – Ecarts au modèle monopolaire en représentation 1 pour $f = 800$ Hz.

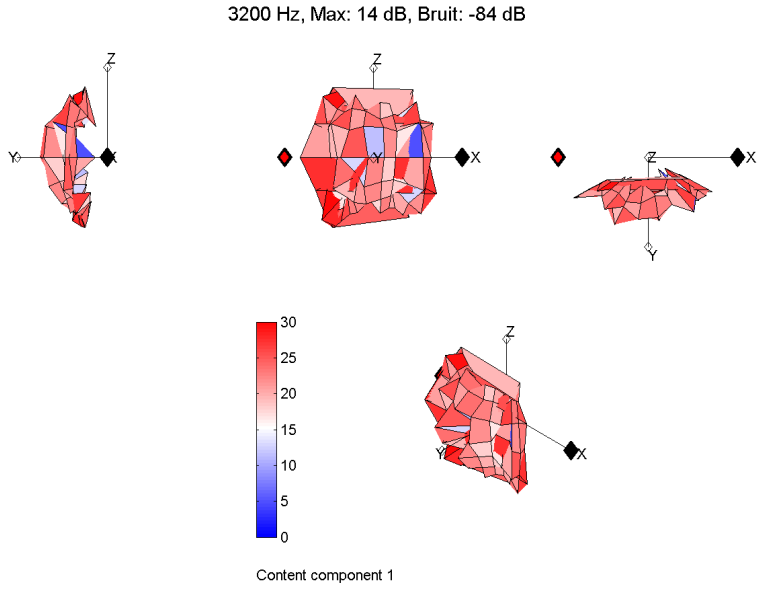



FIG. A.14 – Ecarts au modèle monopolaire en représentation 1 pour $f = 3200$ Hz.

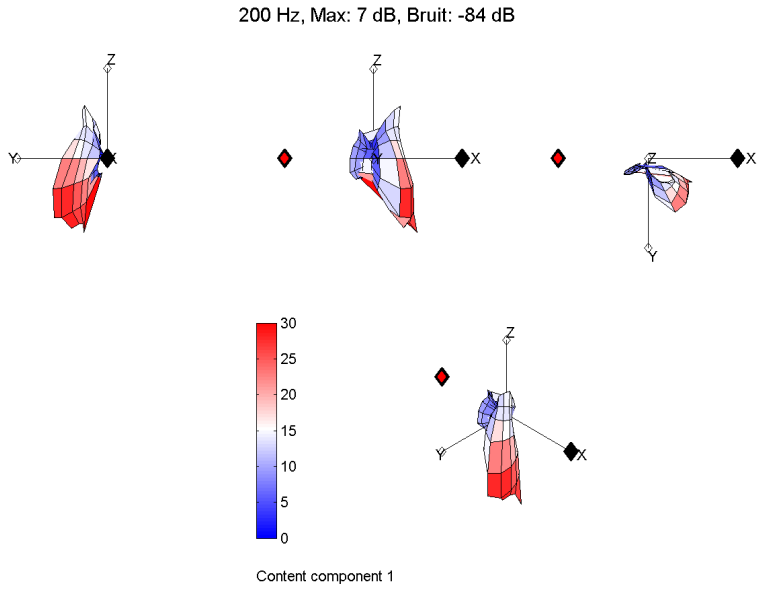



FIG. A.15 – Répartition spatiale du TDH pour $f = 200$ Hz.

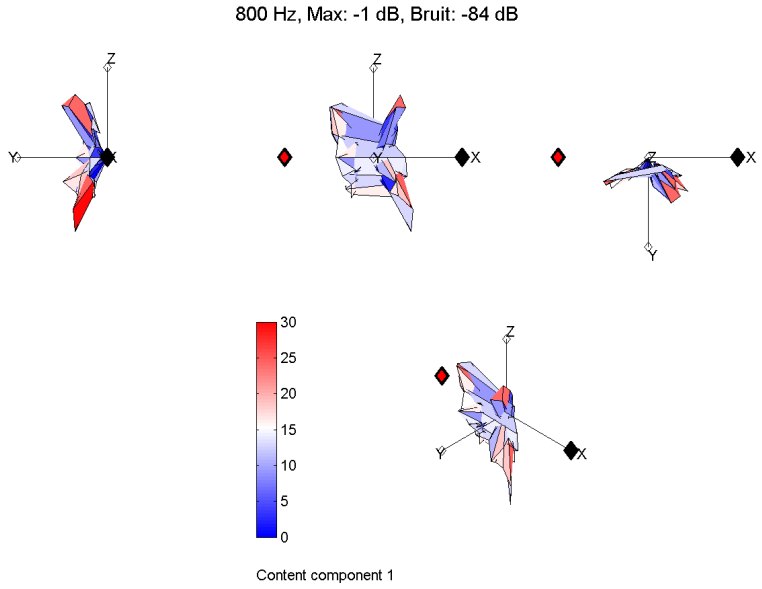



FIG. A.16 – Répartition spatiale du TDH pour $f = 800$ Hz.

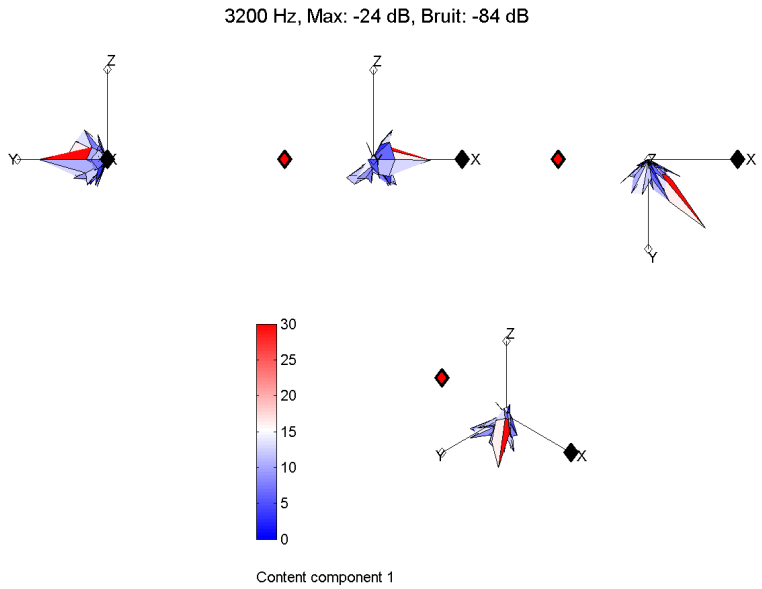



FIG. A.17 – Répartition spatiale du TDH pour $f = 3200$ Hz.

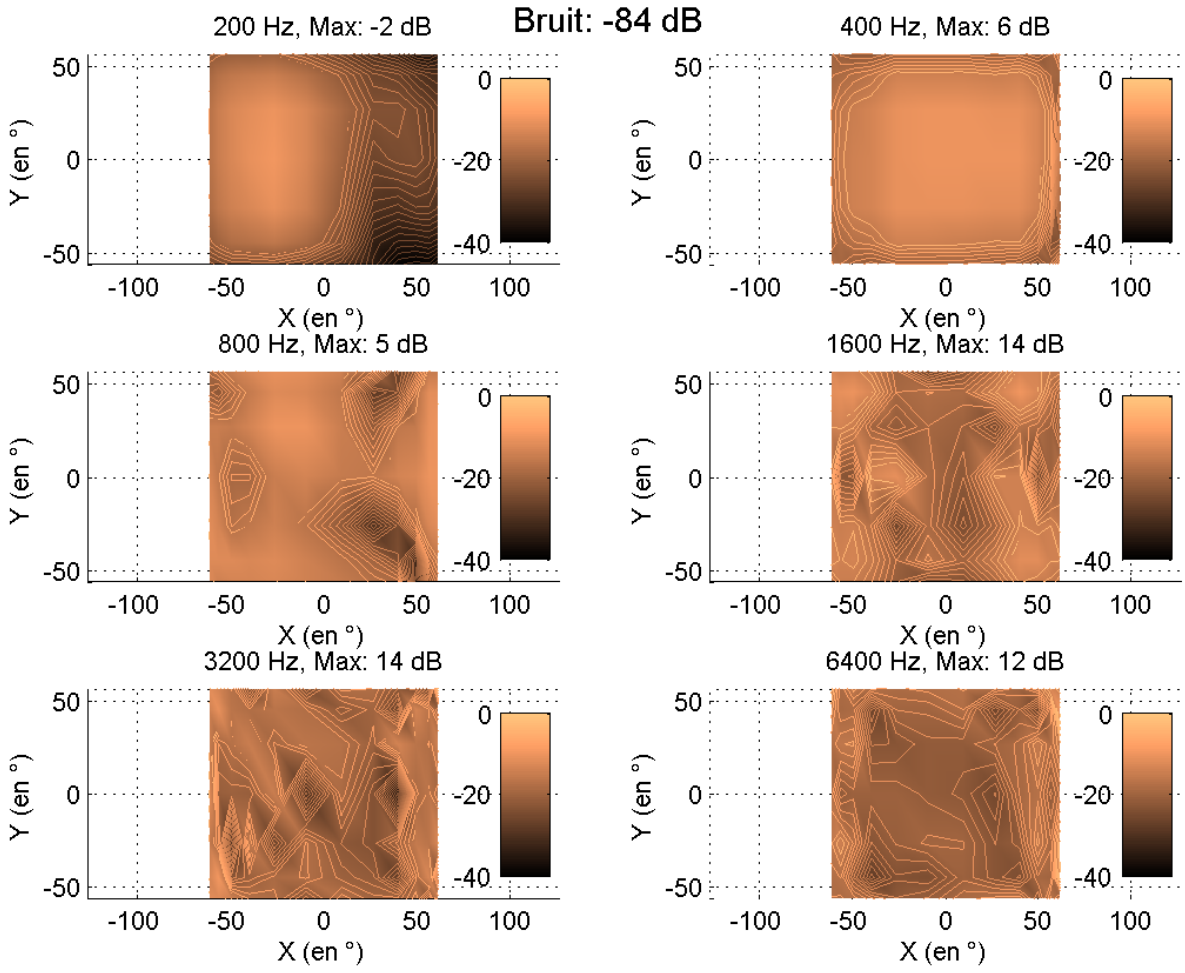



FIG. A.18 – Ecarts au modèle monopolaire en représentation 2.

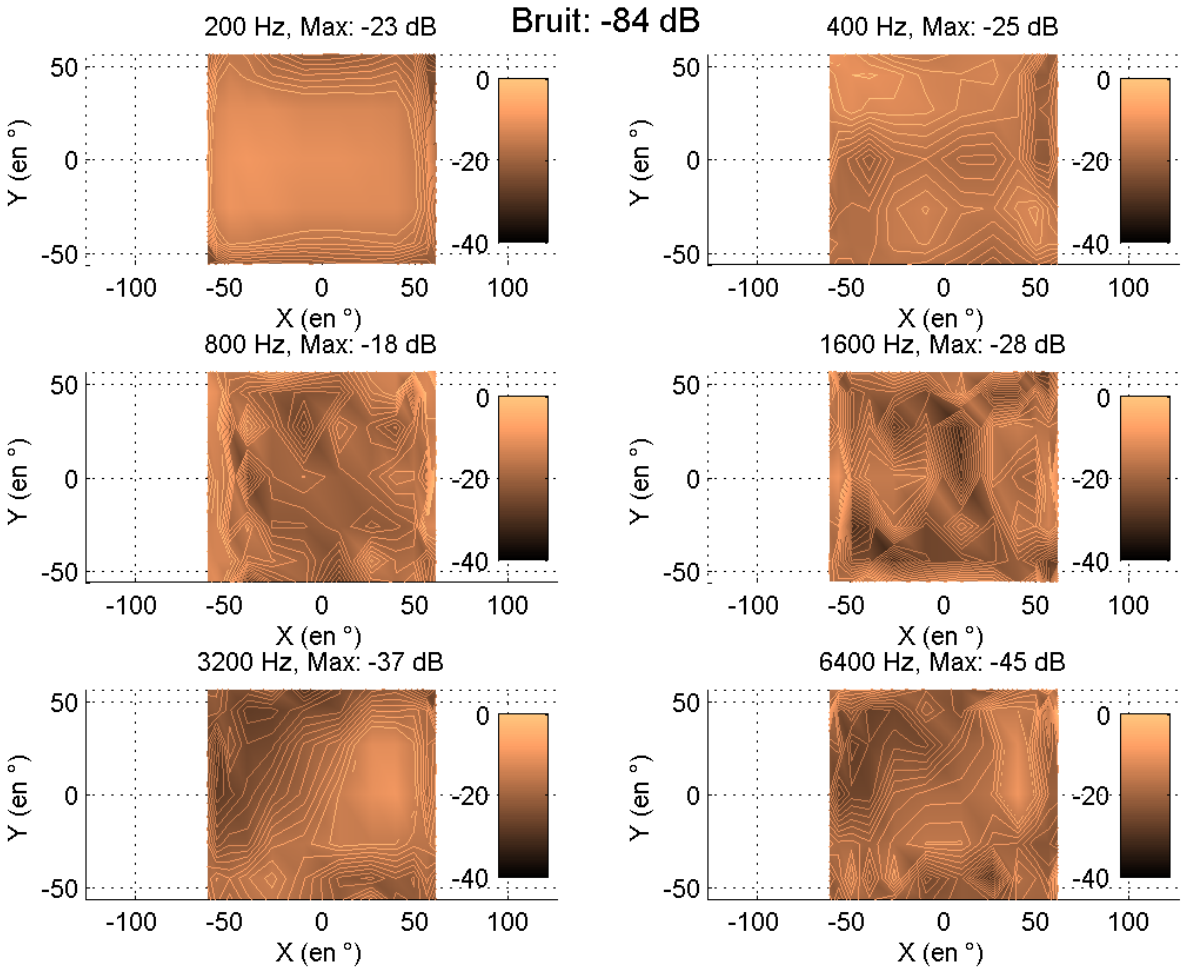



FIG. A.19 – Taux de distorsion harmonique en représentation 2.

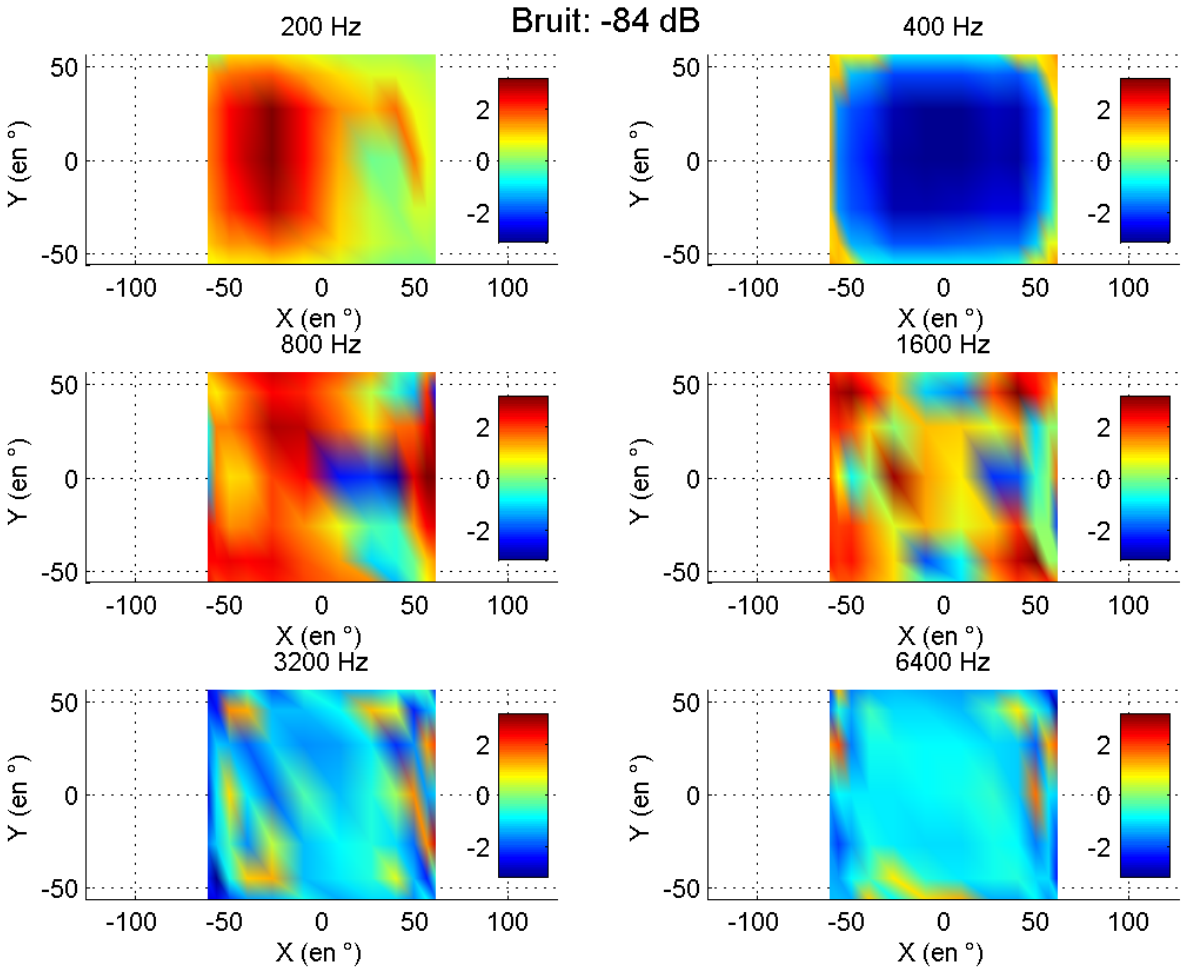



FIG. A.20 – Ecarts au modèle monopolaire en représentation 3.

## A.3 MAP avec encastrement partiel

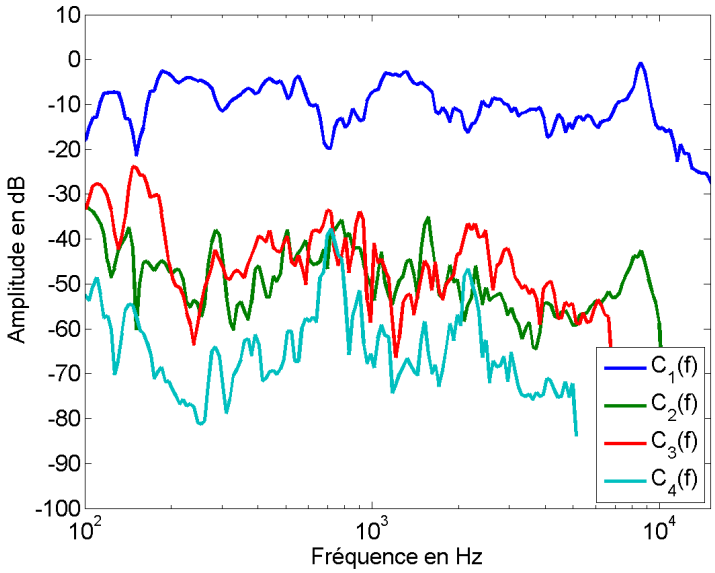



FIG. A.21 – Réponses fréquentielles dans l'axe.

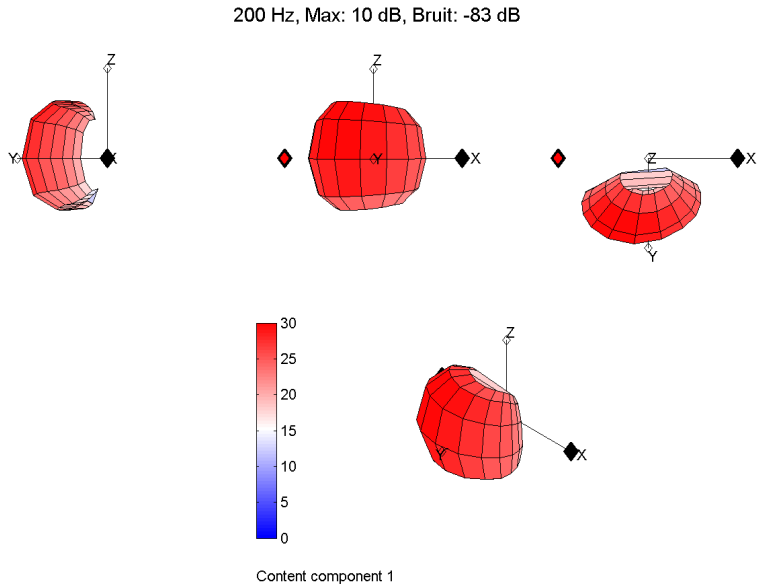



FIG. A.22 – Ecarts au modèle monopolaire en représentation 1 pour $f = 200$ Hz.

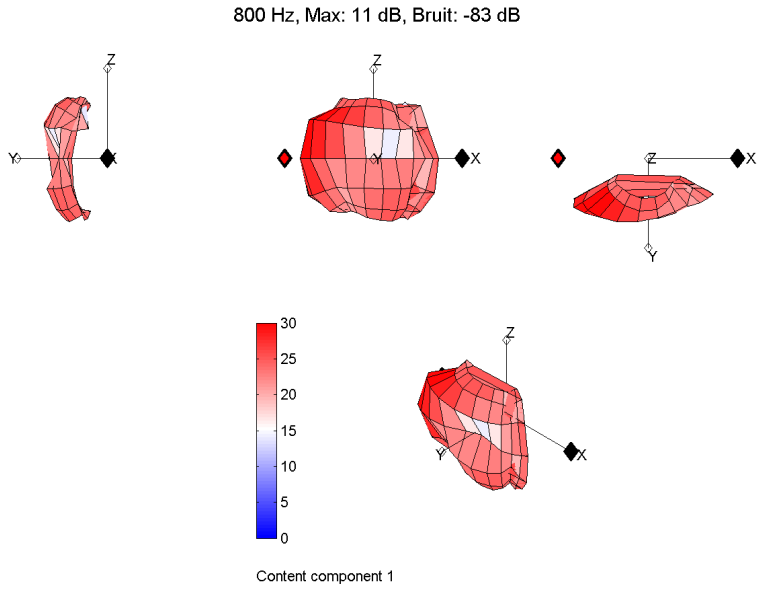



FIG. A.23 – Ecarts au modèle monopolaire en représentation 1 pour $f = 800$ Hz.

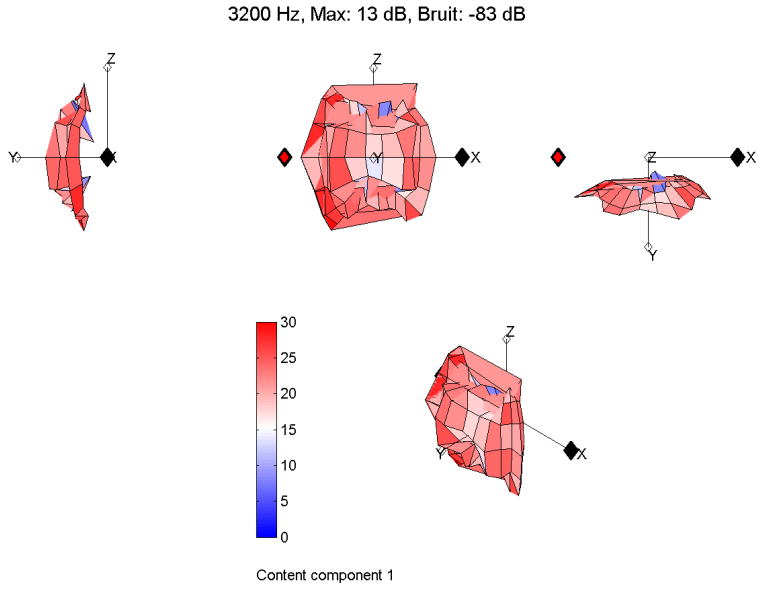



FIG. A.24 – Ecarts au modèle monopolaire en représentation 1 pour $f = 3200$ Hz.

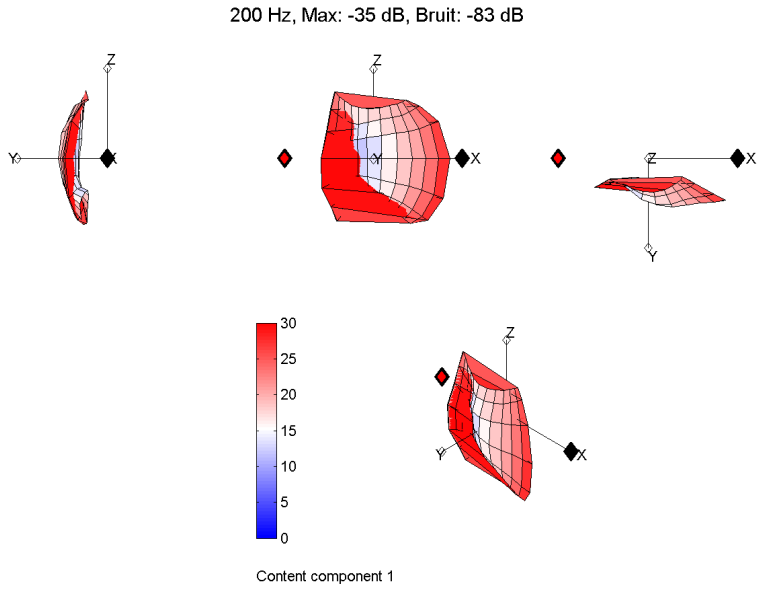



FIG. A.25 – Répartition spatiale du TDH pour $f = 200$ Hz.

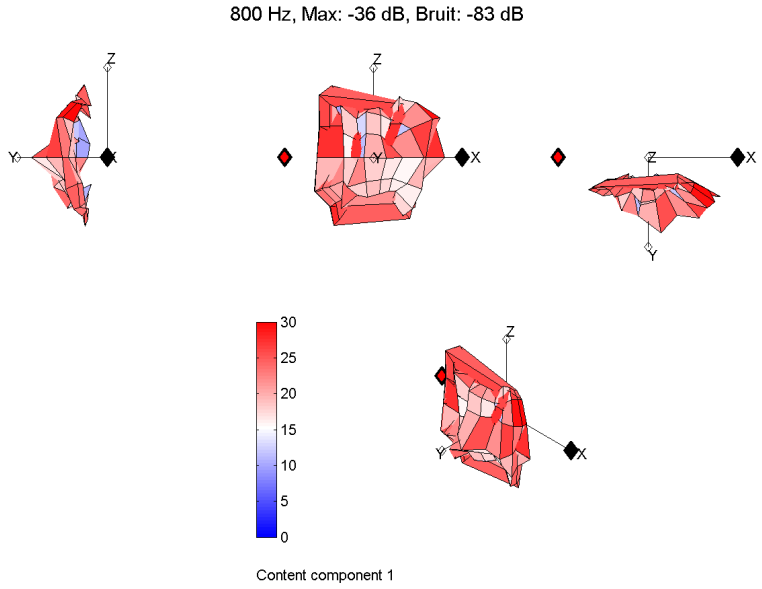



FIG. A.26 – Répartition spatiale du TDH pour $f = 800$ Hz.

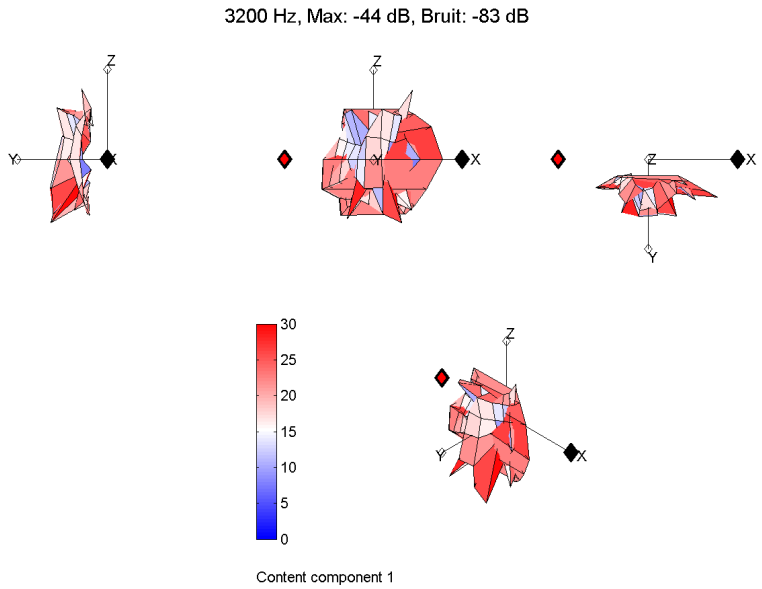



FIG. A.27 – Répartition spatiale du TDH pour $f = 3200$ Hz.

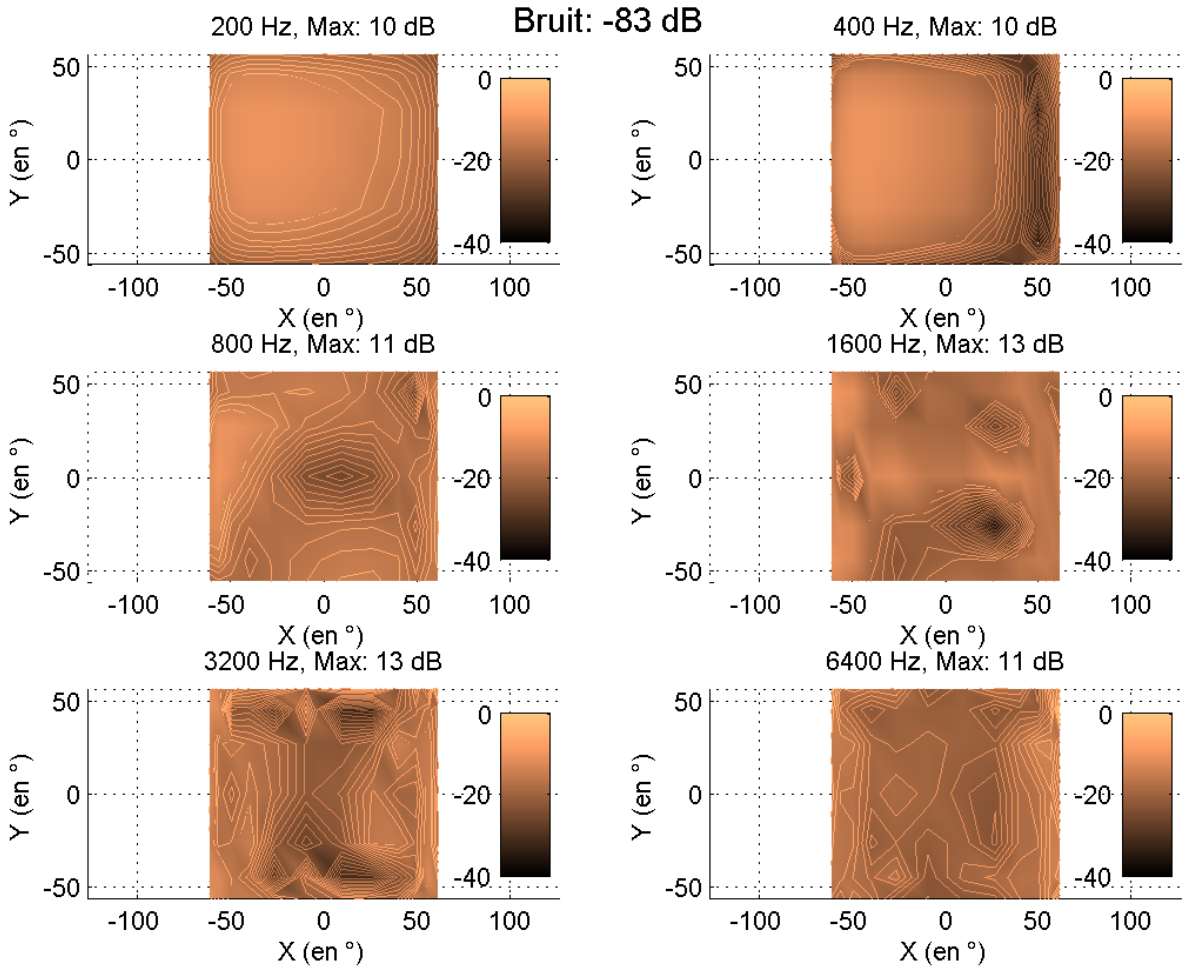



FIG. A.28 – Ecarts au modèle monopolaire en représentation 2.

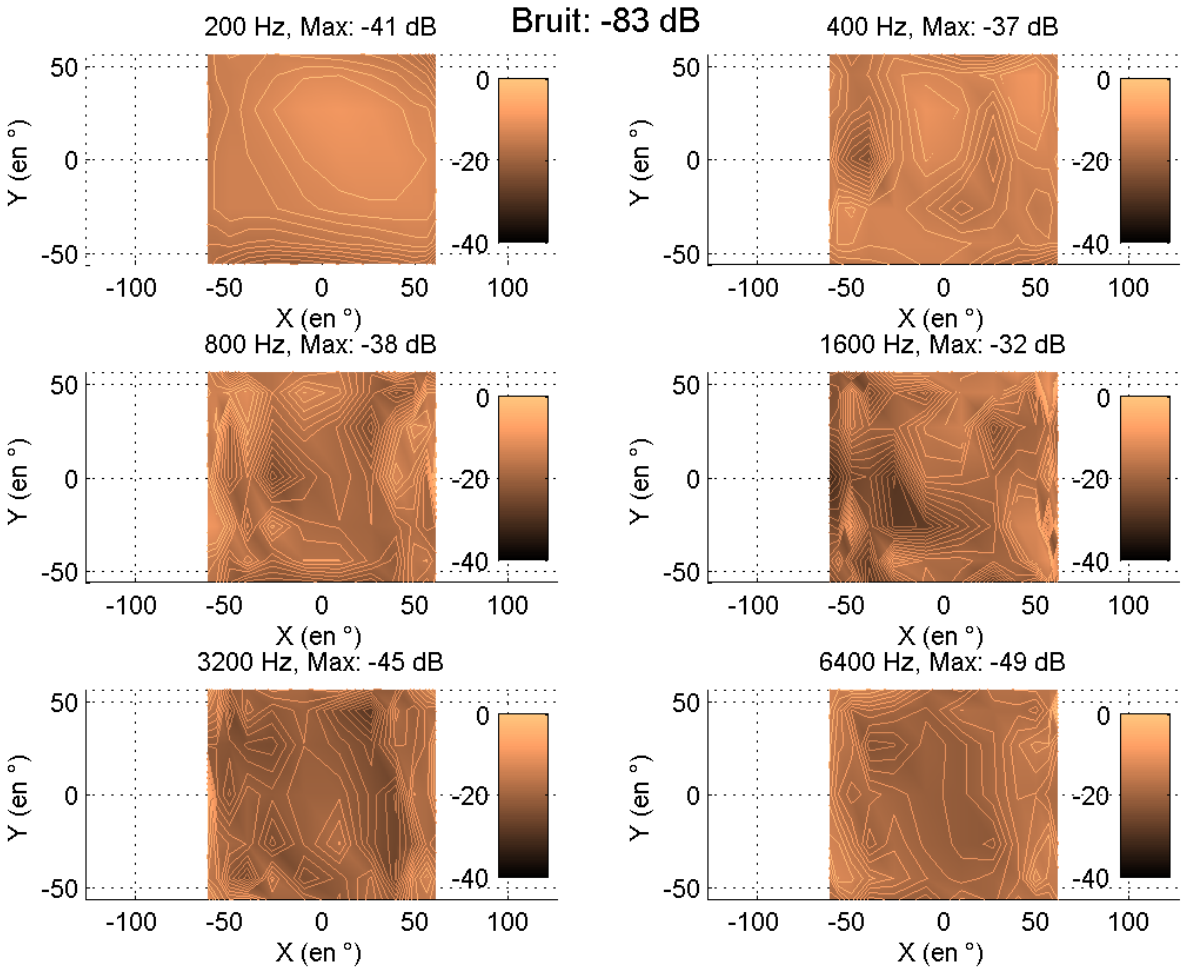



FIG. A.29 – Taux de distorsion harmonique en représentation 2.

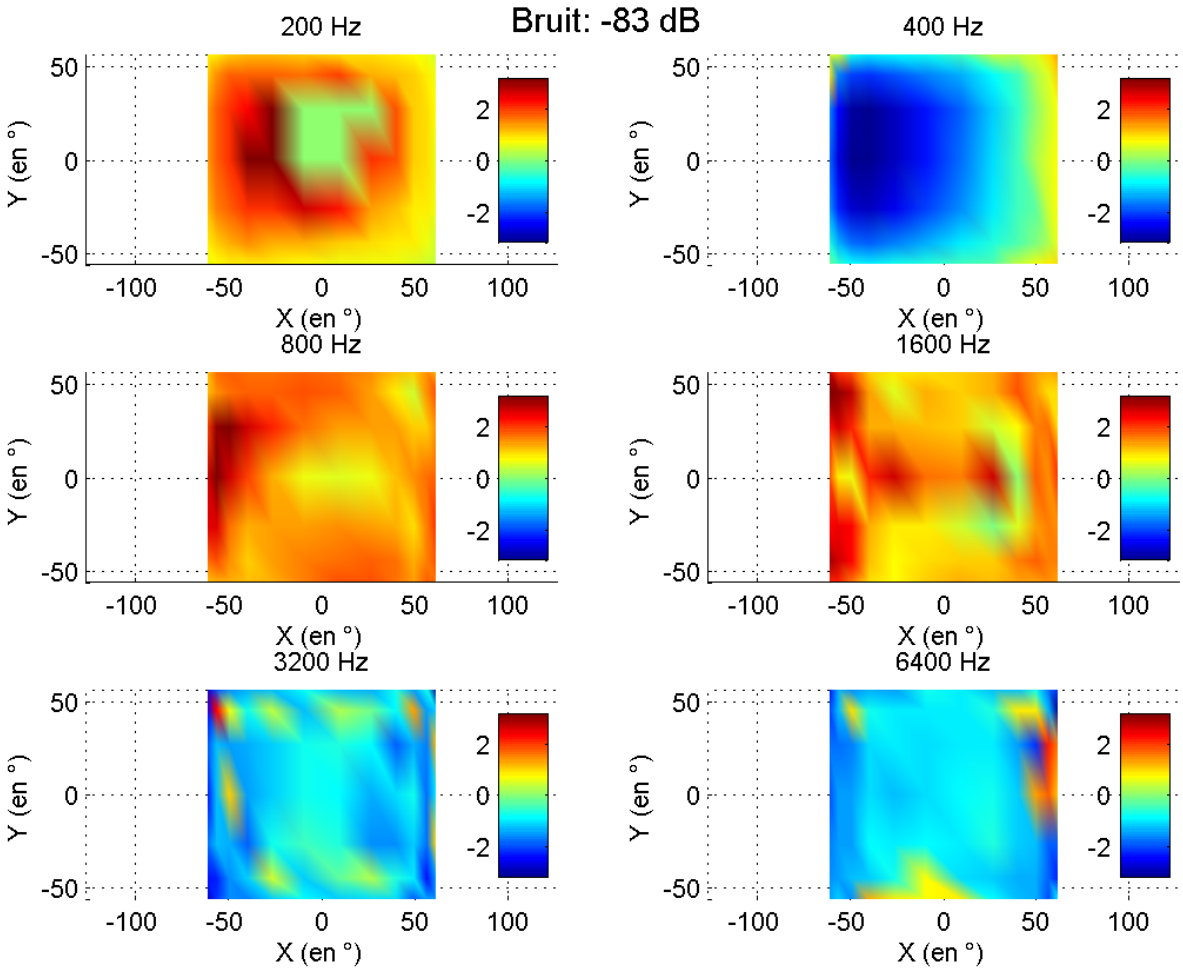



FIG. A.30 – Ecarts au modèle monopolaire en représentation 3.

## A.4 MAP avec encastrement total

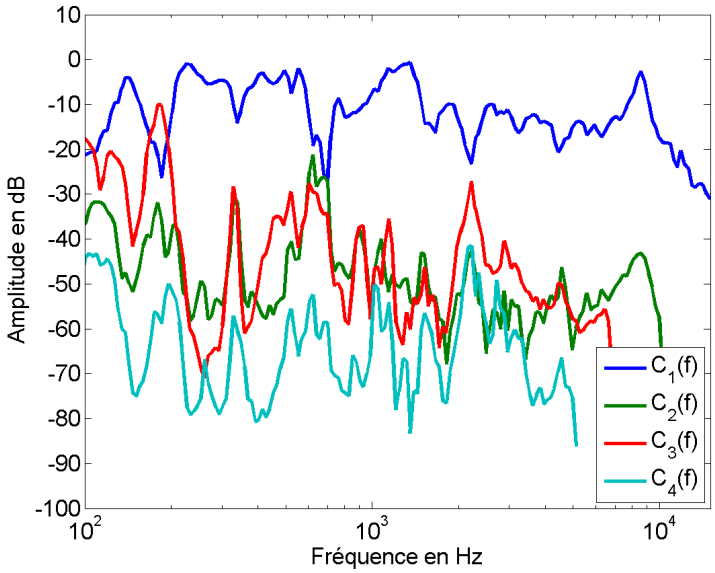



FIG. A.31 – Réponses fréquentielles dans l'axe.

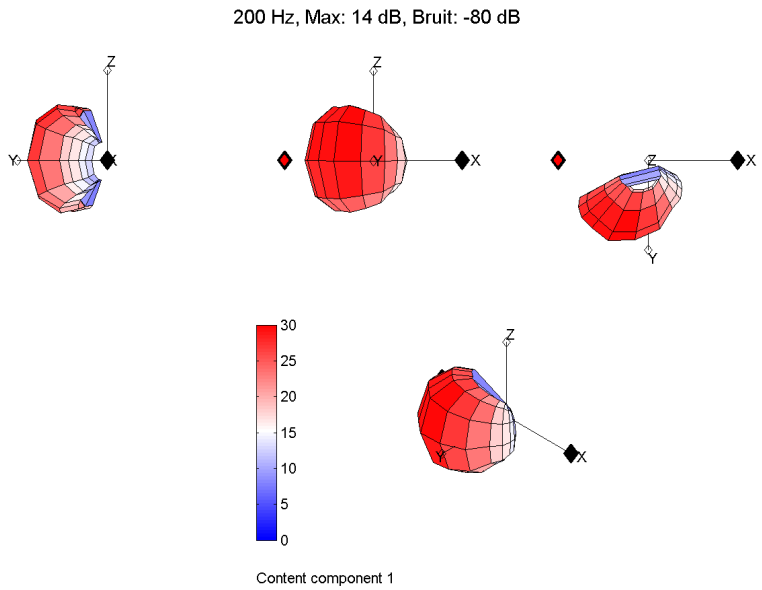



FIG. A.32 – Ecarts au modèle monopolaire en représentation 1 pour $f = 200$ Hz.

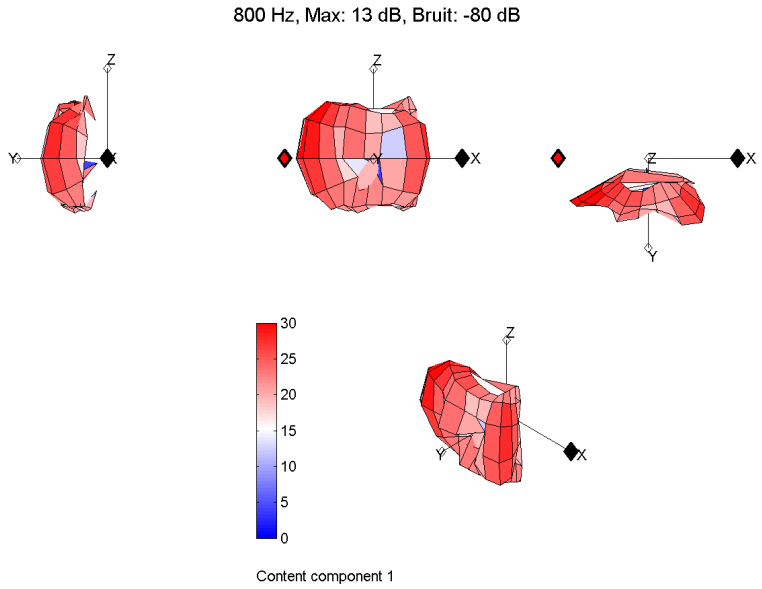



FIG. A.33 – Ecarts au modèle monopolaire en représentation 1 pour $f = 800$ Hz.

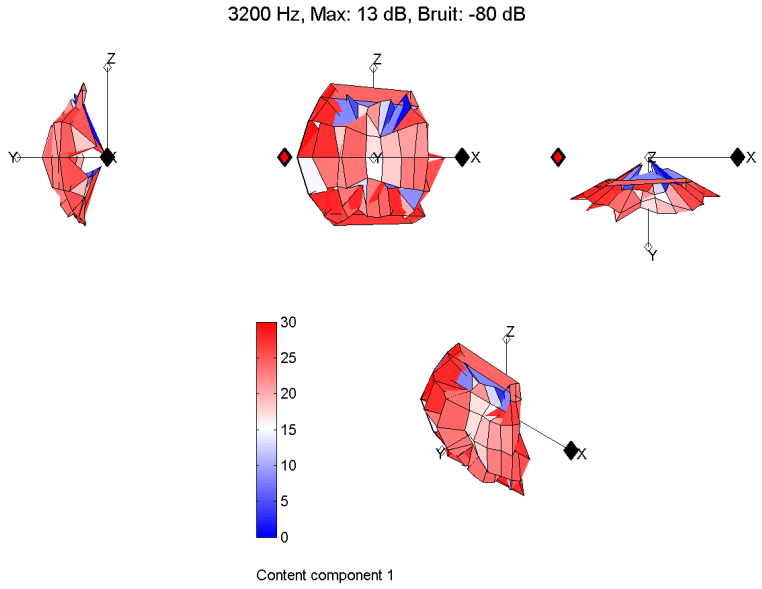



FIG. A.34 – Ecarts au modèle monopolaire en représentation 1 pour $f = 3200$ Hz.

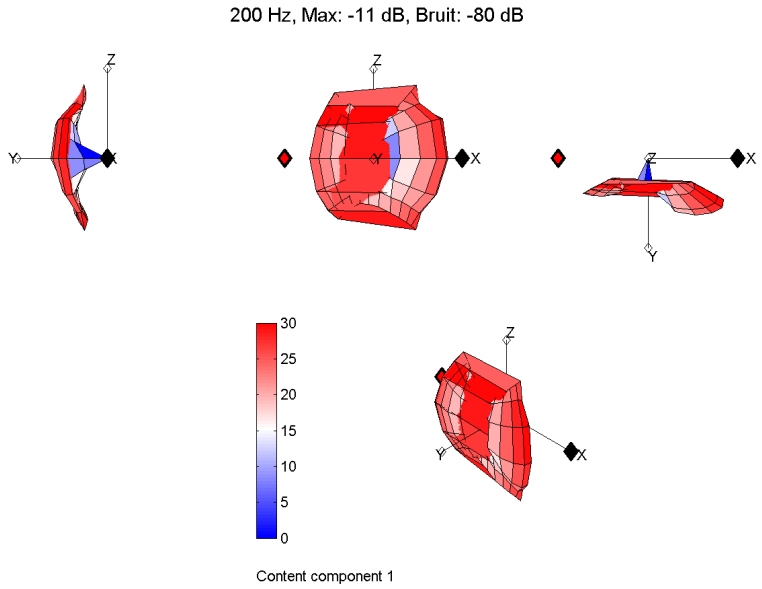



FIG. A.35 – Répartition spatiale du TDH pour $f = 200$ Hz.

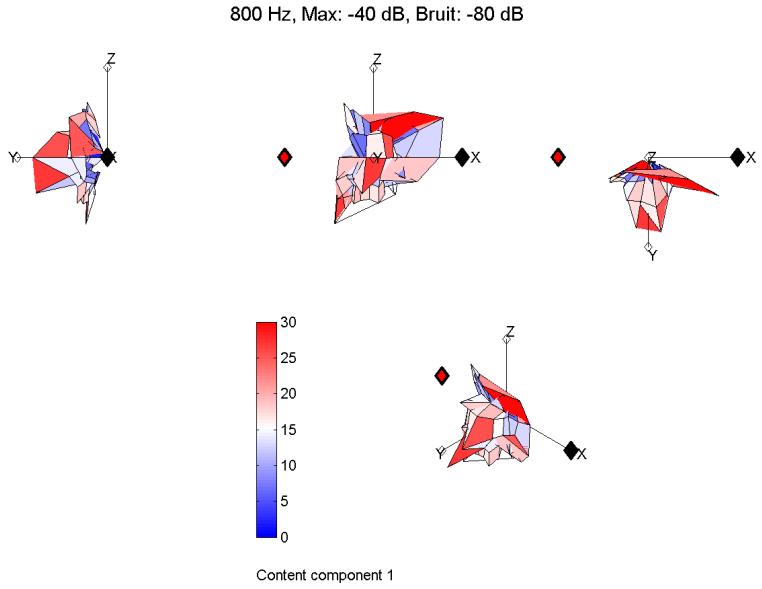



FIG. A.36 – Répartition spatiale du TDH pour $f = 800$ Hz.

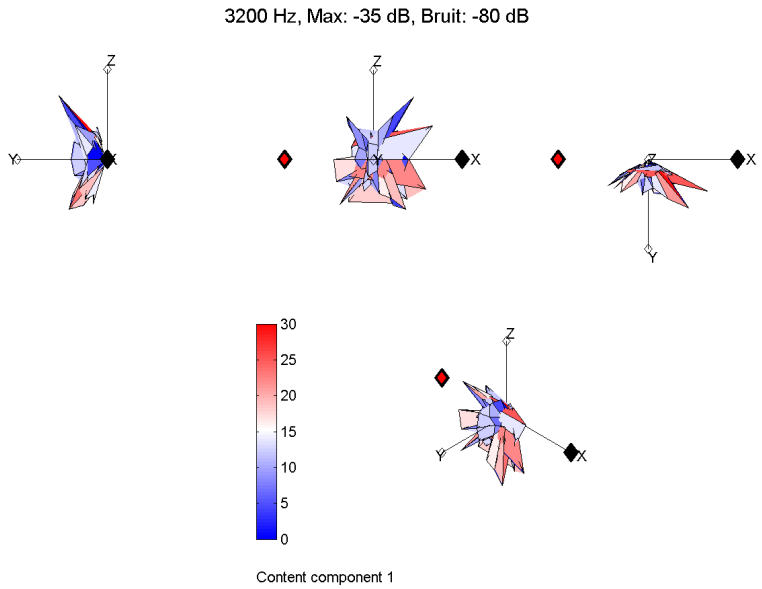



FIG. A.37 – Répartition spatiale du TDH pour $f = 3200$ Hz.

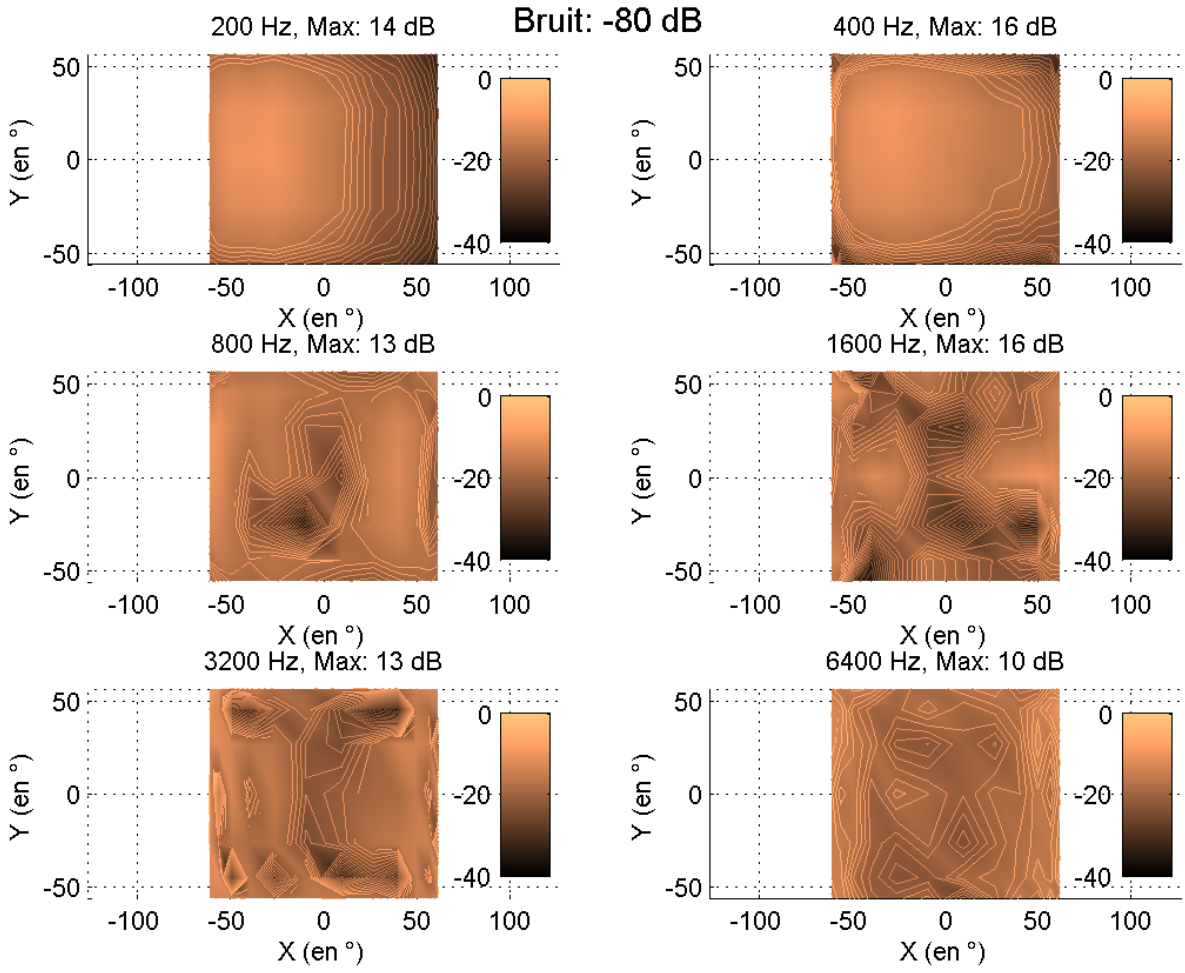



FIG. A.38 – Ecarts au modèle monopolaire en représentation 2.

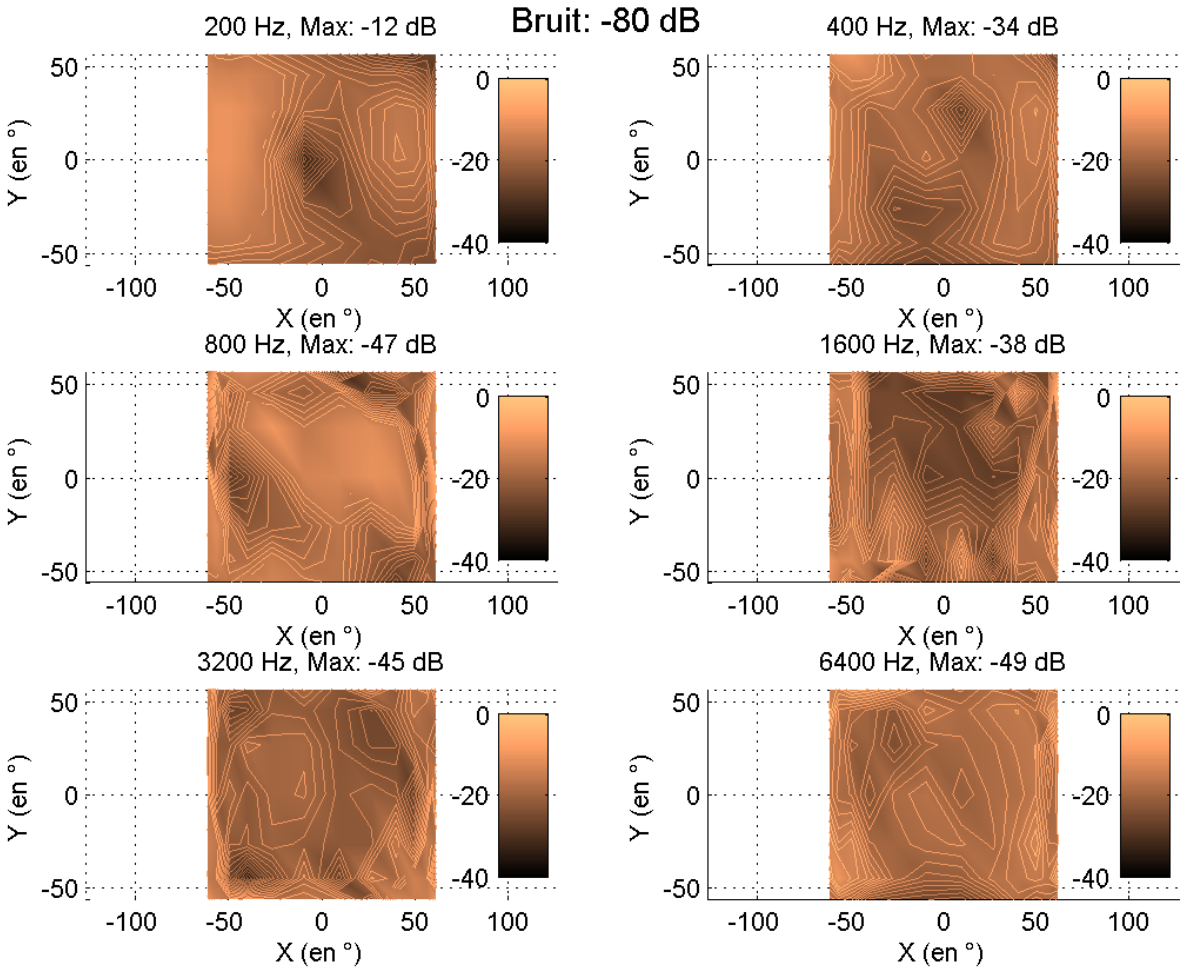



FIG. A.39 – Taux de distorsion harmonique en représentation 2.

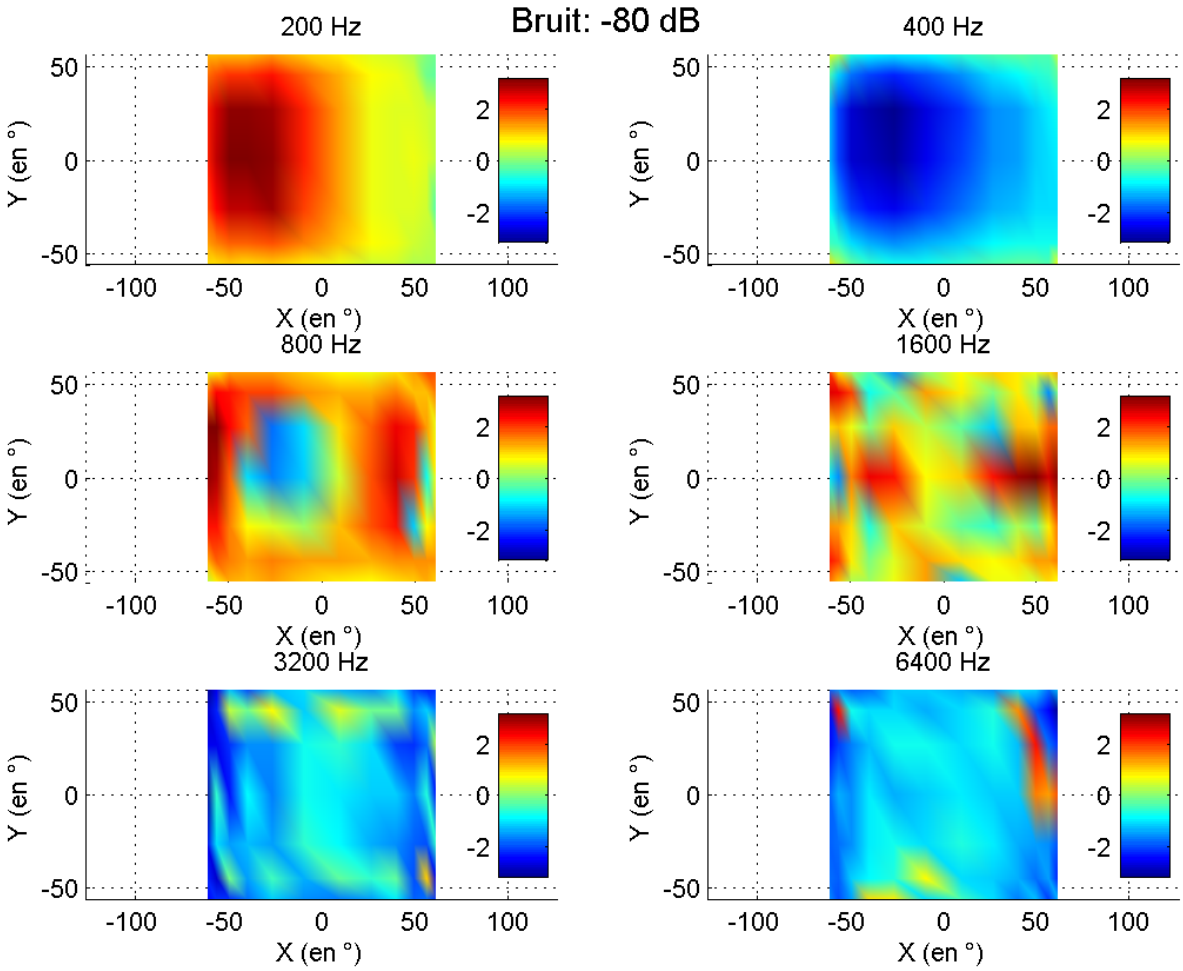



FIG. A.40 – Ecarts au modèle monopolaire en représentation 3.

# Annexe B

# Tracking et cohérence entre les mondes virtuel et physique

Dans cette annexe sont présentés des détails replatifs à l'implémentation du tracking et à la calibration de l'image pour qu'elle soit physiquement cohérente avec le son qui y est associé et avec l'espace physique.

## B.1 Définition d'un repère commun et données géométriques

Pour réaliser une calibration efficace et pour que l'écran soit perçu comme une fenêtre visuelle sur la scène virtuelle il est nécéssaire d'avoir un repère commun, aux mondes réel et virtuel et au trackeur, ainsi qu'un certain nombre de données géométriques.

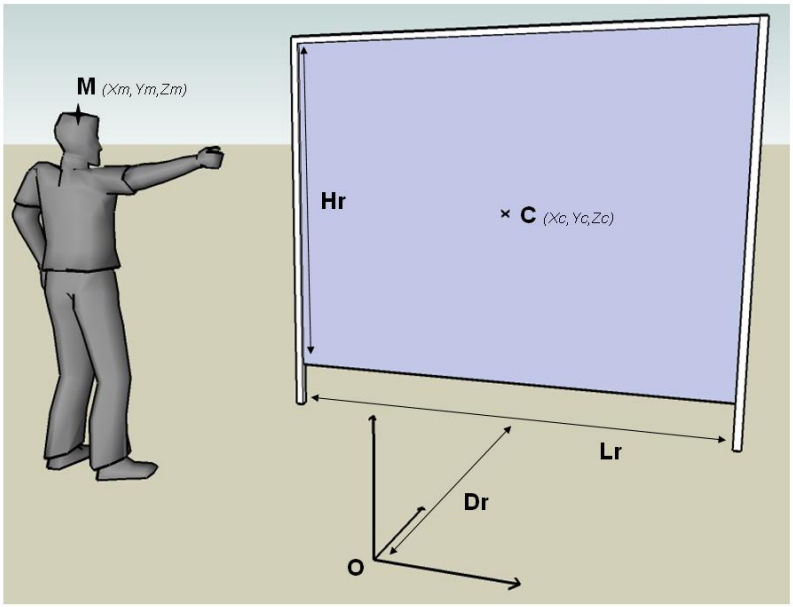



FIG. B.1 – Géométrie de l'installation

## B.2 Création d'une scène VirChor et rendu Open GL

Les scènes virtuelles sont créées dans VirChor, et le choix est fait de considérer les unités arbitraires de VirChor comme équivalentes au mètre pour pouvoir facilement se référer au monde physique.

Le rendu Open GL de la scène 3D ainsi modélisé est réalisé de la façon suivante :

1. Une caméra virtuelle est placée au point d'observation dans la scène virtuelle.
2. Un plan de projection est défini à proximité de la caméra et tous les objets inclus dans ce que l'on appelle le volume de vue (et qui correspond qualitativement à l'angle de vue d'un appareil photo) sont projetés sur ce plan.
3. L'image formée sur ce plan est ensuite renvoyée sur l'écran ou sur le vidéo-projecteur.

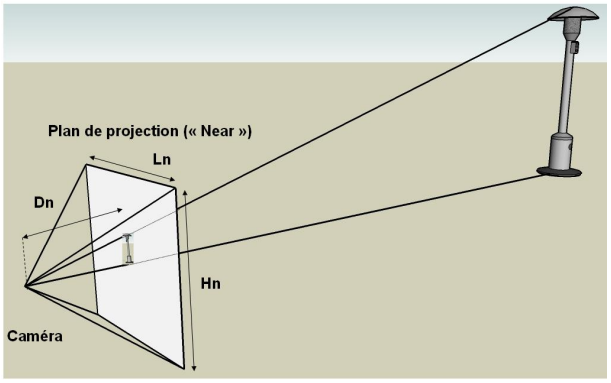



FIG. B.2 – Principe du rendu 3D Open GL

D'un point de vue logiciel, nous disposons de 5 paramètres de réglage pour le plan de projection virtuel :

- $y_{min}$ et $y_{max}$ qui définissent $L_n$ et le décalage longitudinal du plan de projection virtuel par rapport à la caméra (le plan de projection virtuel étant habituellement centré).
- $x_{min}$ et $x_{max}$ qui définissent $H_n$ et le décalage vertical du plan de projection virtuel par rapport à l'utilisateur.
- $Dn$ ou $near$) qui défini la distance entre la caméra et sa projection orthogonale sur le plan de projection virtuel.

## B.3 Conservation des dimensions par passage de l'espace virtuel à l'espace réel.

La démarche est la suivante, si l'on raisonne dans un plan uniquement :

1. Un objet virtuel de taille $L_v$ est disposé dans la scène virtuelle à une distance $D_v$ de l'utilisateur.
2. Cet objet virtuel est projeté sur le plan de projection virtuel et donc sa projection est de taille $L_{near}$.
3. Le plan de projection virtuel est projeté sur l'écran réel, il y a donc une homothétie verticale de rapport $\frac{Hr}{Hn}$ et une homothétie horizontale de rapport $\frac{Lr}{Ln}$ qui sont réalisées. L'objet sur l'écran de restitution est donc maintenant de taille $L_e$.
4. L'utilisateur, à une distance $D_u$ de l'écran, perçoit la distance $D_p$ et la taille $L_p$ de l'objet.

L'objectif de l'opération est que l'angle sous lequel l'objet est perçu (donc la distance à laquelle est l'objet et sa taille) soit conservé, soit :

$$\frac{L_p}{D_p} = \frac{L_v}{D_v}$$

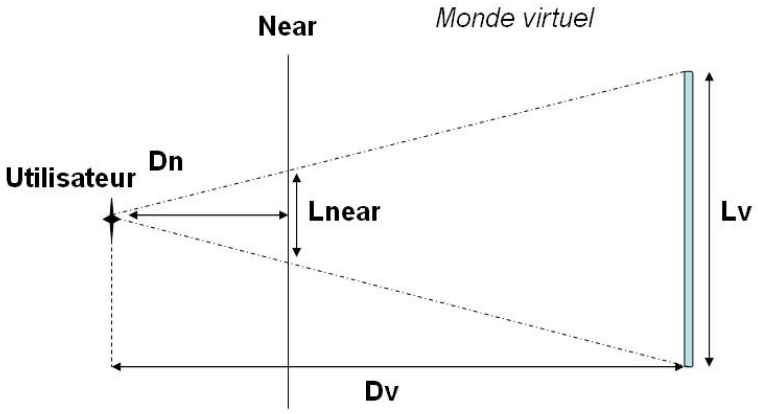



FIG. B.3 – Projection de l'objet virtuel sur le plan de projection virtuel.

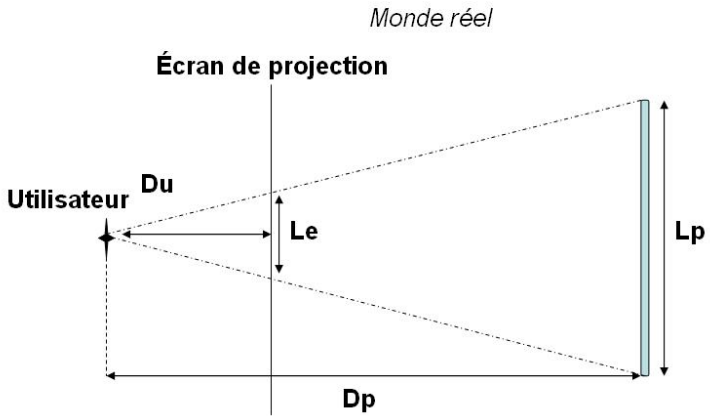



FIG. B.4 – Projection du plan de projection virtuel dans le monde réel et perception de l'image par l'utilisateur.

Alors, on peut écrire, avec le théorème de Thales, dans les mondes virtuel et réel :

$$L_{near} = L_v \frac{Dn}{Dv}$$

$$L_p = L_e \frac{Dp}{Du}$$

L'homothétie à la projection nous donne la relation supplémentaire, écrite ici seulement pour le plan verticale :

$$L_e = L_{near} \frac{Hr}{Hn}$$

Pour que les proportions soient respectées, il faut respecter la condition énoncée précédement, ce qui nous permet, avec $Hn$, $Ln$, $Hr$ et $Lr$ fixés, et en mettant à jour $Du$ avec le trackeur, de déterminer en temps réel les dimensions à donner à $Dn$ pour que les distances soient respectées.

$$Dn = Du\frac{Hn}{Hr} = Du\frac{Ln}{Lr}$$

Les dimensions verticales ($Hr$ et $Hn$) et horizontales ($Lr$ et $Ln$) sont complètement équivalentes de ce point de vue, car elle sont toujours reliées entre elle par un rapport de 4/3.

## B.4 Translation du plan de projection virtuel pour conserver une cohérence visuelle spatiale.

Pour l'instant les distances sont conservées, mais le plan de projection virtuel est toujours centré devant l'utilisateur. L'écran doit être en quelque sorte une fenêtre sur le monde virtuelle au travers de laquelle le point de vue change quand on se déplace dans la zone de rstitution, il faut translater la fenêtre virtuelle relativement aux mouvements de l'utilisateur pour que l'agencement spatial de la scène reste cohérent.

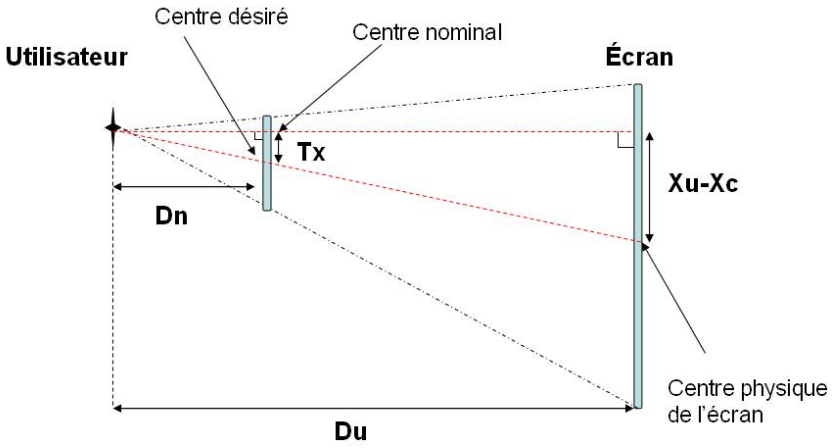



FIG. B.5 – Translation du centre du plan de projection virtuel.

Le théorème de Thalès nous donne les translations horizontales et verticales à réaliser sur le plan de projection virtuel pour avoir une cohérence lorsque l'on se déplace.

$$\text{Verticalement} : T_y = D_n\frac{Y_u - Y_c}{D_u}$$

$$\text{Horizontalement} : T_x = D_n\frac{X_u - X_c}{D_u}$$

Alors, en prenant en compte ces modifications géométriques, en les appliquant en fonction des données envoyées par le trackeur, aux paramètres de rendu OpenGL, un tracking cohérent avec le monde physique est réalisé.