Bases de données pour la séparation de sources

Introduction

Dans ce chapitre, nous présentons deux bases de données que nous avons développées dans le cadre de cette thèse pour des utilisations différentes :

1. une base de données de parole mesurée dans des conditions acoustiques différentes, cette base nous sert à évaluer nos algorithmes de séparation de sources et a été mesurée dans deux milieux acoustiques différents ;
2. une base de données de fonctions de transfert de tête (HRTF) utilisée pour construire les filtres de formation de voies.

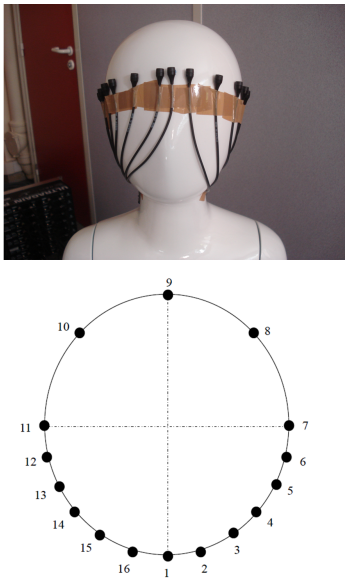
Ces bases de données ont été mesurées avec un réseau de capteurs modélisant le futur robot comme nous le détaillerons dans la première section de ce chapitre. Ces bases de données sont basées sur le calcul des *réponses impulsionnelles* que nous présenterons dans la deuxième section. Le processus d'acquisition sera détaillé dans la dernière section.

7.1 Le réseau de capteurs pour les mesures

Dans le cadre du projet ROMEO, les microphones sont placés sur la tête de l'humanoïde. Dans un premier temps, nous avons modélisé le futur humanoïde par un mannequin de vitrine de taille 1m20 que nous appelons *Theo* (*cf.* figure 7.1a). Nous avons placé 16 capteurs autour de la tête de Théo comme indiqué dans la figure 7.1b. Theo nous a servi à faire les premières mesures en attendant la conception et

la réception du prototype de la tête et torse du robot final Romeo. Pour les mesures, Theo a été mis sur une table tournante, la taille totale du dispositif "Theo + table tournante" est de 1m40 et correspond à la taille de Romeo. Les bases de données enregistrées avec Theo ont pour nom Theo-<nom-de-la-base-de-données>.

(a) Theo, le mannequin de vitrine utilisé pour modéliser le futur robot Romeo



(b) La position des capteurs autour de la tête de Théo (vu de dessus)

Figure 7.1 – Le réseau de capteurs de Theo

7.2 Réponse impulsionnelle acoustique et temps de réverbération

La construction des signaux mélangés et le calcul des HRTF passent par l'estimation des réponses impulsionnelles acoustiques entre différents point de la salle et les capteurs. La *réponse impulsionnelle acoustique* d'un point d'émission vers un capteur caractérise le chemin acoustique entre ces deux points. Elle contient les critères acoustiques de la salle dans laquelle la mesure est faite, typiquement le taux de réverbération de la pièce. Observée dans le domaine temporel, la réponse impulsionnelle acoustique montre les réflexions importantes : le trajet direct, les réflexions précoces et le champ diffus ou réverbéré. L'amplitude maximale de la réponse impulsionnelle correspond à la première onde arrivée, nous pouvons aussi mesurer la durée du trajet direct ainsi que le temps d'arrivée des réflexions précoces. La figure 7.3

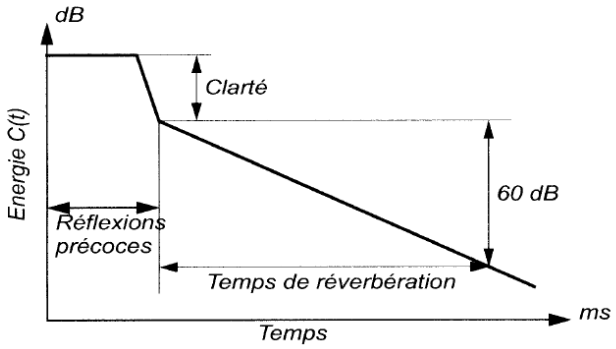


FIGURE 7.2 – Paramètres extraits de la courbe de décroissance de l'énergie (EDC) (cette figure est extraite du manuel de cours d'Yves Grenier [36])

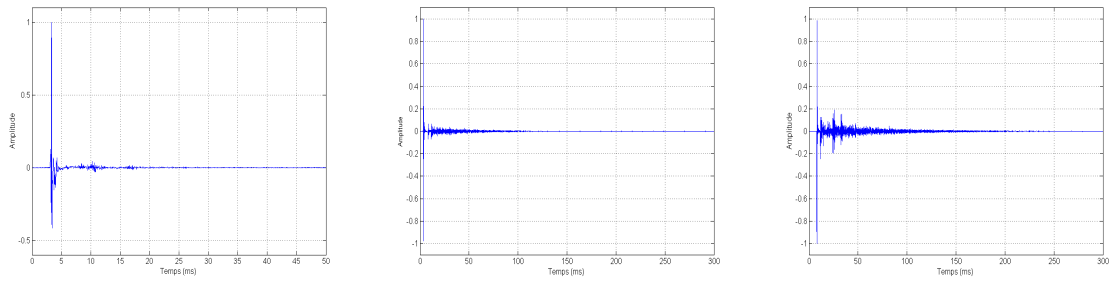
montre différentes réponses impulsionnelles enregistrées dans des conditions acoustiques différentes :

- la figure 7.3a représente une réponse impulsionnelle enregistrée dans une chambre sourde, nous pouvons y distinguer le trajet direct et l'absence quasi-totale de réflexions précoces et de champ diffus ;
- la figure 7.3b représente une réponse impulsionnelle enregistrée dans le studio de Télécom ParisTech ; cette réponse montre que le studio est d'une réverbération modérée, vu l'existence de réflexions précoces avec une amplitude pas très élevée ; le champ diffus quant à lui s'estompe après 100ms ;
- la figure 7.3c est l'estimation de la réponse impulsionnelle de l'Institut de la Vision, nous pouvons distinguer des réflexions précoces d'amplitude plus importante que celle du studio ainsi qu'un champ diffus qui s'estompe plus tardivement ; nous pouvons conclure en regardant ces réponses impulsionnelles que la pièce de l'Institut de la Vision est plus réverbérante que le studio de Télécom ParisTech.

Le calcul du temps de réverbération peut se faire à partir de la *courbe de décroissance de l'énergie* (EDC : Energy Decay Curve). Cette courbe nous permet d'accéder à la durée des réflexions précoces, à la clarté et au temps de réverbération. La courbe de décroissance de l'énergie $C(t)$ est définie comme l'énergie de la réponse impulsionnelle $h(t)$ depuis l'instant t jusqu'à la fin de la réponse, théoriquement $t \to \infty$. Son expression est définie en décibels comme suit :

$$C(t) = 10 \log \sum_{\tau=t}^{\infty} h^2(\tau) \tag{7.1}$$

La figure 7.2 montre l'allure typique de la courbe de décroissance de l'énergie. Le temps d'arrivée du trajet direct est caractérisé par la durée du plateau horizontal. Ensuite, entre le moment où le trajet direct est éliminé de la réponse et les réflexions précoces, la courbe chute brutalement. Ensuite, la courbe EDC décroit régulièrement ce qui correspond à la décroissance exponentielle de la partie de la réponse correspondante au champ diffus ou réverbéré. L'estimation du temps de réverbération se fait à partir de la courbe EDC comme le montre la figure 7.2. Le temps de réverbération est défini comme l'intervalle de temps durant lequel la pression acoustique d'une salle diminue à un millième de sa valeur de régime établi, suite à l'arrêt de la source sonore. Cela représente une diminution du niveau sonore de 60dB et dans ce cas, le temps de réverbération est noté RT_{60}. Dans un environnement réel, il est difficile d'obtenir une décroissance de 60dB du niveau sonore, il est donc plus commun d'utiliser RT_{30} ou RT_{20} qui représentent le temps que prend une source pour décroitre de 30dB ou de 20dB respectivement. La figure 7.4 montre les courbes de décroissance de l'énergie obtenues dans la chambre anéchoïque, dans le studio de Télécom ParisTech et à l'Institut De la Vision.

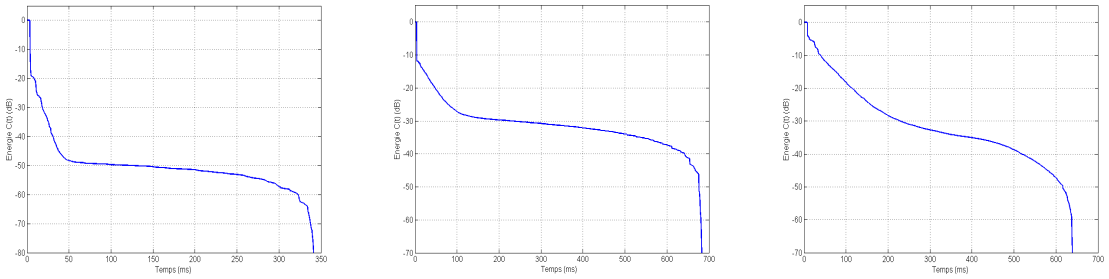


(a) Chambre anéchoïque (b) Studio Télécom ParisTech (c) Institut de la vision

Figure 7.3 – Réponse impulsionnelle entre un point dans une pièce et le microphone 1

7.3 Construction des mélanges convolutifs

Une bonne évaluation des algorithmes de séparation de sources nécessite, à égale importance, des outils d'évaluation des performances de séparation qui donnent une description complète de la *qualité* des signaux séparés (*cf.* chapitre 8 section 8) et une *bonne* base de données d'évaluation. Une base de données d'évaluation des algorithmes de séparation de sources doit être :



(a) Chambre anéchoïque (b) Studio Télécom ParisTech (c) Institut de la vision

FIGURE 7.4 – Courbes de décroissance de l'énergie (EDC) entre un point dans une pièce et le premier microphone

- assez large pour permettre une évaluation des performances de séparation statistiquement significative ;
- enregistrée dans des conditions acoustiques différentes allant de la chambre anéchoïque à la pièce avec une réverbération significative et ceci afin de tester la robustesse des algorithmes de séparation face à la réverbération ;
- variée afin de tester la capacité des algorithmes de séparation à généraliser. Pour une base de données de parole, elle doit contenir des extraits de parole enregistrés avec des voies différentes d'hommes et de femmes, les sources audio doivent provenir de directions d'arrivées différentes et variées.

Enregistrer directement avec le réseau de capteurs des voix d'hommes et de femmes dans des conditions acoustiques différentes et des situations variées prend beaucoup de temps et nécessite un grand nombre de locuteurs volontaires et un nombre plus grand de mesures.

Pour construire nos bases de données de parole dans des conditions réelles sans avoir à répéter les enregistrement avec les locuteurs volontaires, nous considérons d'abord une base de données de parole enregistrée dans des conditions anéchoïques, c'est à dire sans réverbération ni bruit, nous appelons cette base de données Sources-pures. Les mélanges convolutifs sont construits comme suit :

1. dans chaque condition acoustique (pièce), estimer les réponses impulsionnelles acoustique $h_{ij}(l)$ de différents points d'émission $1 \leq i \leq N$ (différentes directions d'arrivées) vers chacun des microphones du réseau de capteurs $1 \leq j \leq M$;
2. calculer les contributions aux capteurs $[s_{ij}(t)]_{1\leq i\leq N,\, 1\leq j\leq M}$ pour chacune des sources
   $[s_i(t)]_{1\leq i\leq N}$ considérées et des points d'émission voulu ; pour chaque capteur j

et chaque source i, la contribution $s_{ij}(t)$ est la convolution d'une source pure $s_i(t)$ avec la réponse impulsionnelle acoustique d'un point d'émission choisi et le capteur j, ceci donne : $s_{ij}(t) = \sum_{l=0}^{L-1} h_{ij}(l)\, s_i(t-l)$;

3. pour chaque capteur, le signal reçu consiste en la somme des N sources images calculées au niveaux de ce capteur : $x_j(t) = \sum_{i=0}^{N} s_{ij}(t)$.

Pour chaque pièce, les réponses impulsionnelles acoustiques sont calculées une seule fois pour chacun des points sources considérés vers chaque capteur, nous pouvons ensuite changer de locuteur hors enregistrement et autant de fois que l'on veut en utilisant la base de données des sources pures.

7.4 Calcul des réponses impulsionnelles avec les séquences complémentaires de Golay

Nous considérons un système à temps discret caractérisé par sa réponse impulsionnelle $h(t)$, un signal d'entrée $s(t)$ et un signal de sortie $x(t)$, $t \in \mathbb{Z}$. Pour identifier le système, nous avons besoin d'estimer $h(t)$ pour un signal d'entrée $s(t)$ connu et un signal de sortie $x(t)$. Dans notre cas, le système représente le chemin acoustique entre un point d'émission et un capteur et nous voulons estimer la réponse impulsionnelle acoustique $h(t)$:

$$x(t) = s(t) * h(t) \tag{7.2}$$

où $*$ est un opérateur de convolution.

Pour estimer cette réponse impulsionnelle, nous utilisons les séquences complémentaires de Golay [33] comme signal d'entrée. Les séquences complémentaires de Golay ont la propriété intéressante que leur fonctions d'autocorrélation ont des lobes secondaires complémentaires : la somme des fonctions d'autocorrélation est nulle partout sauf à l'origine. Les séquences complémentaires de Golay ne sont pas uniques [34]. Nous utilisons une paire de séquences $a(t)$ et $b(t)$ de longueur L et définie comme suit :

$$\begin{aligned} a(t) &= \pm 1 \text{ pour } 1 \leq t \leq L \\ b(t) &= \pm 1 \text{ pour } 1 \leq t \leq L \end{aligned} \tag{7.3}$$

Ces séquences sont des séquences complémentaires de Golay si et seulement si :

$$a(-t) * a(t) + b(-t) * b(t) = 2L\delta(t) \tag{7.4}$$

où $\delta(t)$ est l'impulsion de Dirac. Les séquences que nous utilisons sont définies récursivement comme suit :

$$\begin{bmatrix} A_L \\ B_L \end{bmatrix} = \begin{bmatrix} A_{L/2} & B_{L/2} \\ A_{L/2} & -B_{L/2} \end{bmatrix} \text{ avec } \begin{matrix} A_L = [a(1) \cdots a(L)] \\ B_L = [b(1) \cdots b(L)] \end{matrix}$$

$$\text{et } \begin{bmatrix} A_2 \\ B_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

Les réponses du système aux entrées $a(t)$ et $b(t)$ sont respectivement :

$$\begin{aligned} x_a(t) &= a(t) * h(t) \\ x_b(t) &= b(t) * h(t) \end{aligned} \tag{7.5}$$

En utilisant les équations (7.4) et (7.5), la réponse impulsionnelle du système est donnée par :

$$h(t) = \frac{1}{2L}\left(a(-t) * x_a(t) + b(-t) * x_b(t)\right) \tag{7.6}$$

Dans un cas pratique, nous utilisons comme entrée une séquence de Golay construite comme le montre la figure 7.5. Chaque séquence de Golay A et B de longueur L est répétée N_G fois. Après la mesure, le premier bloc de chacune des deux séries qui sont répétées n'est pas pris en compte dans l'estimation de la réponse impulsionnelle, car ce premier bloc contient une convolution avec le bloc précédent, qui pour la première série était un bloc nul, et pour la seconde série était un bloc de la première série. Pour le reste des blocs mesurés $x_A = [x_a(1) \ldots x_a(L)]$ et $x_B = [x_b(1) \ldots x_b(L)]$, nous estimons les réponses impulsionnelles dans le domaine fréquentiel. La réponse impulsionnelle finale est la moyenne des réponses impulsionnelles estimées. Cette répétition assure une certaine robustesse dans la mesure de la réponse.

Cette méthode de calcul des réponses impulsionnelles a été utilisée pour le calcul des réponses impulsionnelles acoustiques dans différentes pièces réverbérantes et pour l'estimation des réponses impulsionnelles de tête (HRIR : Head Related Transfer Functions) utilisées pour calculer les fonctions de transfert de tête comme nous l'avons présenté dans la section 3.4. Dans la suite, nous présenterons le détail de ces mesures.

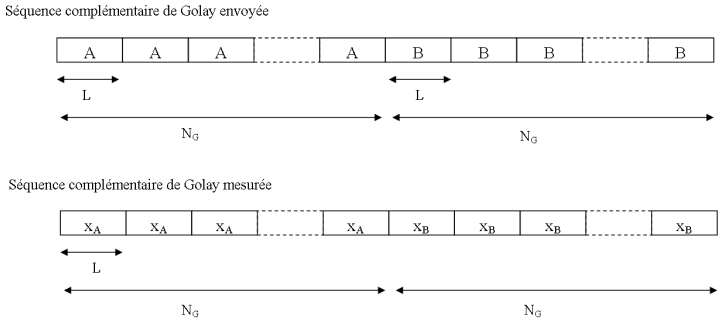


Figure 7.5 – Structure totale des séquences complémentaires de Golay à l'entrée et celles mesurées

7.5 Base de données des réponses impulsionnelles dans un milieu réverbérant

7.5.1 A Télécom ParisTech

Nous avons enregistré une première base de réponses impulsionnelles dans le studio de Télécom ParisTech dont le taux de réverbération est de RT_{60}= 300 ms, calculé à partir de la courbe de décroissance d'énergie 7.4b. Nous avons estimé les réponses impulsionnelles de différentes positions comme le montre la figure 7.6. Les points d'émission sont placés à 1m20 du réseau de capteurs et nous avons mesuré les réponses impulsionnelles de -90° à 90° avec un pas de 10° (à l'exception de -10° et 10° où nous n'avons pas fait de mesure).

Nous avons choisi d'évaluer les algorithmes sur un cas de séparation de 2 et 3 sources. Pour un cas de séparation de deux sources, la première source est placée à 0° et la seconde source est choisie entre 20° et 90°. Pour un cas de séparation de trois sources, la première source est toujours placée à 0° et la $2^{ème}$ et $3^{ème}$ source sont choisies entre -90° et -20° et entre 20° et 90°. Nous appelons ces bases de données Theo-RI-studio.

7.5.2 A l'institut de la vision

Dans le cadre du projet Romeo, un appartement témoin a été conçu à l'institut de la vision. Cet appartement est une vitrine pour le projet Romeo, c'est un appartement témoin qui sert à l'acquisition de base de données commune aux différents partenaires du projet et à faire les premiers tests de Romeo dans un environnement proche de celui dans lequel l'humanoïde évoluera. Dans cet appartement témoin,

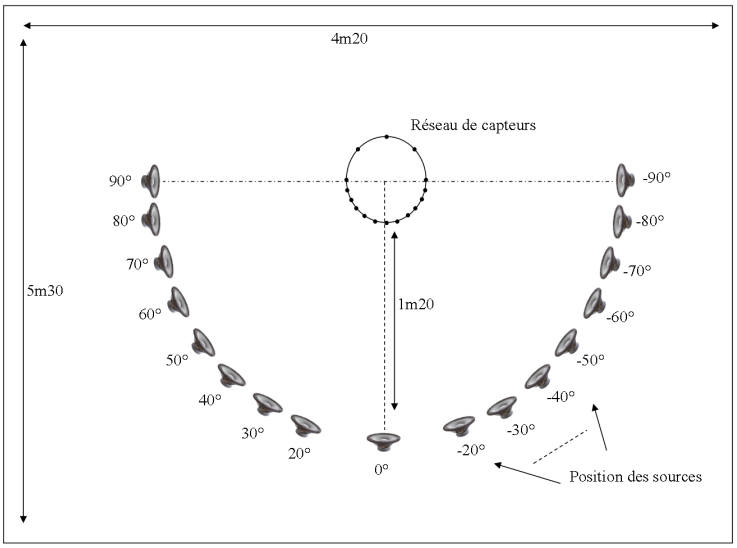


FIGURE 7.6 – Position des sources par rapport au réseau de capteurs dans le studio de Télécom ParisTech

nous avons procédé à des mesures de réponses impulsionnelles de trois angles d'arrivée différents : 20°, 60° et -35° comme montré à la figure 7.7. Comme l'indique la courbe de décroissance de l'énergie 7.4c, le temps de réverbération de cet appartement témoin de l'institut de la vision est de RT_{60}= 600 ms. Nous appelons ces bases de données Theo-RI-IDV.

7.6 Base de données de HRTF

Nous rappelons qu'une fonction de transfert de tête ou HRTF est la représentation fréquentielle d'une réponse impulsionnelle de tête ou HRIR qui caractérise comment un signal émis d'une direction spécifique est reçu à une oreille. La HRIR de chaque oreille capture l'information de localisation du signal source et l'altération produite par la tête et le pavillon sur le champ sonore proche [16]. Nous étendons le principe des HRIR et HRTF à la problématique de l'audition des robots avec un réseau de capteurs, donc avec plus de deux "oreilles", nous mesurons les HRIR et HRTF pour les 16 capteurs de Theo.

A notre connaissance, ceci représente un premier cas d'une base de données de HRTF et HRIR multicapteurs avec différents angles d'azimut et d'élévation ; jusqu'à maintenant, les bases de données disponibles pour le téléchargement sont binaurales [12, 27, 28, 35]. KEMAR est une base de données binaurale mesurée avec une tête

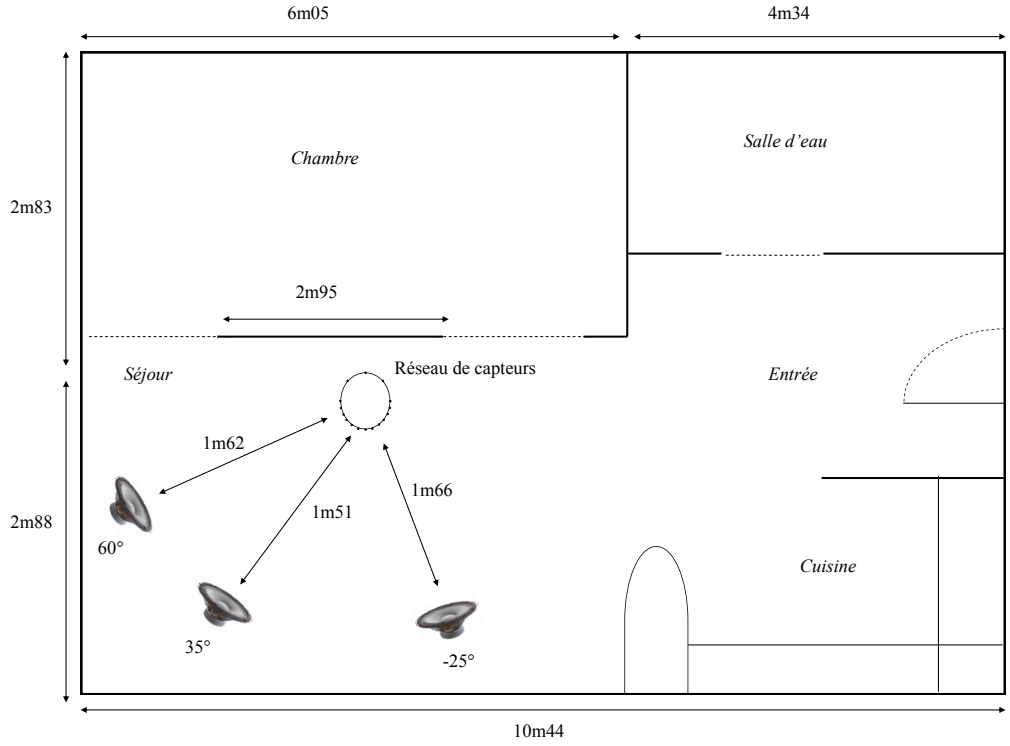


Figure 7.7 – Position des sources par rapport au réseau de capteurs dans l'Institut De la Vision (IDV)

de mannequin pour 70 positions [35] et la base de données de HRTF CIPIC a été mesurée avec 45 sujets humains pour 25 azimuts et 50 élévations [12].

Le calcul des HRIR se fait en utilisant les séquences complémentaires de Golay, dans la suite la description du processus de mesure.

7.6.1 Description matérielle et logicielle

L'estimation des réponses impulsionnelles de tête en utilisant les séquences complémentaires de Golay a été faite dans la salle anéchoïque de Télécom ParisTech comme le montre la figure 7.8. Nous avons utilisé le matériel suivant :

- Theo pour l'acquisition de la base de données des HRIR et HRTF : Theo-HRTF ;
- deux enregistreurs Echo Audiofire Pre8 [1] ;

1. http://www.echoaudio.com/Products/FireWire/AudioFirePre8/index.php

– 16 microphones AKG C417 pp [2] (*cf.* figure 11.2) ;
– 7 haut-parleurs Tannoy System 600 [3] ;
– une table tournante Brüel & Kjær Type 9640 [4].

Les enregistreurs Audiofire Pre8 sont reliés par FireWire à un PC ayant comme système d'exploitation Linux avec un noyau temps réel. Les logiciels utilisés sont :

– Le kit de connexion audio Jack (JACK-control) [46] une application open-source qui contrôle le serveur son spécifique à l'infrastructure Linux Audio Desktop.
– Ffado [32], un pilote open-source pour les dispositifs pro-audio se basant sur des connexions par FireWire. Nous utilisons ffado-mixer pour contrôler la synchronisation entre les enregistreurs et accéder à la table de mixage des canaux des enregistreurs.

7.6.2 Processus expérimental

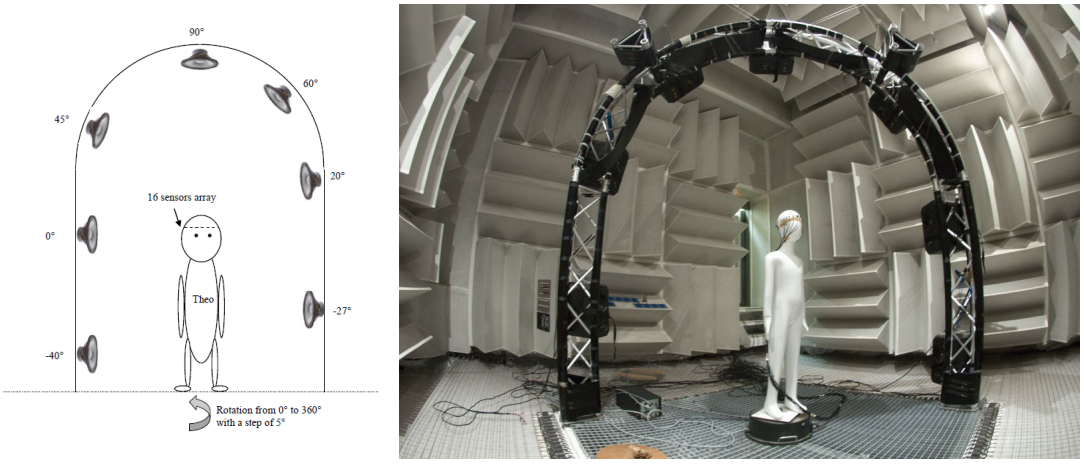


FIGURE 7.8 – Theo et la position des hauts-parleurs pour l'enregistrement des séquences complémentaires de Golay dans la chambre anéchoïque de Télécom ParisTech

Dans la chambre anéchoïque, nous avons mesuré des HRTF depuis 504 points d'azimut et d'élévation distribués comme suit (*cf.* figure 7.8) :

– 73 angles d'azimut de 0° à 355° avec un pas de 5°
– 7 angles d'élévation : -40°, -27°, 0°, 20°, 45°, 60° et 90°

2. http://www.akg.com/mediendatenbank2/psfile/datei/35/c4174055c447d8838.pdf
3. http://www.tannoy.com/products/158/uman_System600.pdf
4. http://www.bksv.com/products/Télécomaudiosolutions/electroacousticsaccessories/turntablesystemtype9640.aspx

Theo est fixé sur une table tournante dans le centre de l'arc supportant les haut-parleurs. Une séquence complémentaire de Golay est émise séquentiellement de chaque haut-parleur (chaque élévation) et enregistrée avec un réseau de capteurs de 16 microphones pour chaque angle d'azimut. Cette base de données Theo-HRTF est disponible avec les fréquences d'échantillonnage de 16kHz et 48kHz et peut être téléchargée par ce lien `http://www.tsi.telecom-paristech.fr/aao/?p=347`

Conclusion

Nous avons présenté dans le chapitre les différentes bases de données que nous avons acquises avec une configuration de réseaux de 16 capteurs placés autour de la tête d'un mannequin de vitrine Theo. La base de données des HRTF sert à calculer les filtres de formation de voies utilisés dans l'étape de prétraitement de nos algorithmes de séparation à deux étapes. La base de données de mélanges convolutifs, acquise dans deux milieux différents, sert à évaluer et comparer les algorithmes de séparation de sources proposés. Le tableau 7.1 présente un récapitulatif de ces bases de données.

Base de données	Nombre de données
Sources-pures	40 paires
Theo-RI-studio	320 paires (de 0° à -90°)
Theo-RI-IDV	120 paires
Theo-HRTF	504 HRIR/HRTF

Table 7.1 – Récapitulatif des bases de données enregistrées avec Theo (16 capteurs chacun)