Évaluer l'apport du binaural dans une application mobile audiovisuelle

Julian Moreira

le cnam

École doctorale Informatique, Télécommunications et Électronique (Paris)

Centre d'Études et de Recherches en Informatique et Communications

THÈSE DE DOCTORAT

présentée par : **Julian Moreira**

soutenue le : **10 juillet 2019**

pour obtenir le grade de : **Docteur du Conservatoire National des Arts et Métiers**

Spécialité : **Informatique**

Évaluer l'apport du binaural dans une application mobile audiovisuelle

THÈSE dirigée par :

M. Natkin Stéphane	*Professeur Émérite,* Cnam
Mme. Viaud-Delmon Isabelle	*Chargée de Recherche CNRS, HDR,* Ircam

ENCADRANTES :

Mme. Gros Laetitia	*Ingénieur de recherche,* Orange Labs
Mme. Nicol Rozenn	*Ingénieur de recherche, HDR,* Orange Labs
Mme. Le Prado Cécile	*Maître de conférences associée,* Cnam

RAPPORTEURS :

M. Paquier Mathieu	*Professeur des universités,* UBO
M. Jouvelot Pierre	*Directeur de recherche,* MINES ParisTech

PRÉSIDENT DU JURY :

M. Garcia Alexandre	*Professeur des universités,* Cnam

Résumé

Les terminaux mobiles offrent à ce jour des performances de plus en plus élevées (CPU, résolution de l'écran, capteurs optiques, etc.) Cela rehausse la qualité vidéo des services média, que ce soit pour le visionnage de contenu vidéo (streaming, TV, etc.) ou pour des applications interactives telles que le jeu vidéo. Mais cette évolution concernant l'image n'est pas ou peu suivie par l'intégration de systèmes de restitution audio de haute qualité dans ce type de terminal. Or, parallèlement à ces évolutions concernant l'image, des solutions de son spatialisé sur casque, à travers notamment la technique de restitution binaurale basée sur l'utilisation de filtres HRTF (*Head Related Transfer Functions*) voient le jour.

Dans ce travail de thèse, nous nous proposons d'évaluer l'intérêt que peut présenter le son binaural lorsqu'il est utilisé sur une application mobile audiovisuelle. Une partie de notre travail a consisté à déterminer les différents sens que l'on pouvait donner au terme « application mobile audiovisuelle » et parmi ces sens ceux qui d'une part étaient pertinents et d'autre part pouvaient donner lieu à une évaluation comparative avec ou sans son binaural.

Le couplage entre son binaural et visuel sur mobile occasionne en premier lieu une question d'ordre perceptive : comment peut-on organiser spatialement une scène virtuelle dont le son peut se déployer tout autour de l'utilisateur, et dont le visuel est restreint à un si petit écran ? La première partie de cette thèse est consacrée à cette question. Nous menons une expérience visant à étudier le découplage spatial possible entre un son binaural et un visuel rendus sur smartphone. Cette expérience révèle une forte tolérance de l'être humain face aux dégradations spatiales pouvant survenir entre les deux modalités. En particulier, l'absence d'individualisation des HRTF, ainsi qu'un très grand découplage en élévation ne semblent pas affecter la perception. Par ailleurs, les sujets semblent envisager la scène « comme si » ils y étaient eux-mêmes directement projetés, à la place de la caméra, et cela indépendamment de leur propre distance à l'écran. Tous ces résultats suggèrent la possibilité d'une association entre son binaural et visuel sur mobile dans des conditions d'utilisation proches du grand public.

Dans la seconde partie de la thèse, nous tentons de répondre à la question de l'apport du binaural en déployant une expérience « hors les murs », dans un contexte plausible d'utilisation grand public. Trente sujets jouent dans leur vie quotidienne à un jeu vidéo de type Inifinite Runner, développé pour l'occasion en deux versions, une avec du son binaural, et l'autre avec du son monophonique. L'expérience dure cinq semaines, à raison de deux sessions par jour. Ce protocole procède de la méthode dite Experience Sampling Method, sur l'état de l'art de laquelle nous nous sommes appuyés. Nous calculons à chaque session des notes d'immersion, de mémorisation et de performance, et nous comparons les notes obtenues entre les deux versions sonores. Les résultats indiquent une immersion significativement meilleure pour le binaural. La mémorisation et la performance ne sont en revanche pas soumises à un effet statistiquement significatif du rendu sonore. Au-delà des résultats, cette expérience nous permet de discuter de la question de la validité des données en fonction de la méthode de déploiement, en confrontant notamment bienfondé théorique et faisabilité pratique.

Mots clés : Binaural, smartphone, qualité d'expérience, expérience utilisateur, perception audiovisuelle, effet ventriloque, point d'alignement spatial subjectif, méthode d'échantillonnage de l'expérience, contexte, attribut sonore, jeu vidéo

Abstract

In recent years, smartphone and tablet global performances have been increased significantly (CPU, screen resolution, webcams, etc.). This can be particularly observed with video quality of mobile media services, such as video streaming applications, or interactive applications (e.g., video games). However, these evolutions barely go with the integration of high quality sound restitution systems. Beside these evolutions though, new technologies related to spatialized sound on headphones have been developed, namely the binaural restitution model, using HRTF (Head Related Transfer Functions) filters.

In this thesis, we study the potential contribution of the binaural technology to enhance the quality of experience of an audiovisual mobile application. A part of our work has been dedicated to define what is an "audiovisual mobile application", what kind of application could be fruitfully experienced with a binaural sound, and among those applications which one could lead to a comparative experiment with and without binaural.

In a first place, the coupling of a binaural sound with a mobile-rendered visual tackles a question related to perception : how to spatially arrange a virtual scene whose sound can be spread all around the user, while its visual is limited to a very small space ? We propose an experiment in these conditions to study how far a sound and a visual can be moved apart without breaking their perceptual fusion. The results reveal a strong tolerance of subjects to spatial discrepancies between the two modalities. Notably, the absence or presence of individualization for the HRTF filters, and a large separation in elevation between sound and visual don't seem to affect the perception. Besides, subjects consider the virtual scene as if they were projected inside, at the camera's position, no matter what distance to the phone they sit. All these results suggest that an association between a binaural sound and a visual on a smartphone could be used by the general public.

In the second part, we address the main question of the thesis, i.e., the contribution of binaural, and we conduct an experiment in a realistic context of use. Thirty subjects play an Infinite Runner video game in their daily lives. The game was developed for the occasion in two versions, a monophonic one and a binaural one. The experiment lasts five weeks, at a rate of two sessions per day, which relates to a protocol known as the "Experience Sampling Method". We collect at each session notes of immersion, memorization and performance, and compare the notes between the monophonic sessions and the binaural ones. Results indicate a significantly better immersion in the binaural sessions. No effect of sound rendering was found for memorization and performance. Beyond the contribution of the binaural, we discuss about the protocol, the validity of the collected data, and oppose theoretical considerations to practical feasibility.

Keywords : Binaural sound, smartphone, quality of experience, user experience, audiovisual perception, ventriloquist effect, point of subjective spatial alignment, experience sampling method, context, sound attribute, video game

Remerciements

Quoi qu'on en dise, le travail d'une thèse est un travail essentiellement solitaire. Pour un doctorant en informatique de surcroît, dont l'activité quotidienne ne s'écarte jamais bien loin de son ordinateur, la plupart des journées s'organisent autour d'activités qui ne souffrent pas la collaboration directe, qu'il s'agisse de la lecture bibliographique, de l'organisation d'une expérience, du dépouillement des résultats, ou de la rédaction du manuscrit. Pourtant, et fort heureusement, le thésard n'est pas seul. De nombreuses personnes croisent sa route, qui lui permettent d'accompagner, de discuter, de valider, voire même de le divertir de son travail. Certains d'entre eux participent de façon ponctuelle et précise, d'autres de façon plus régulière ou diffuse. Tous méritent de recevoir les remerciements qui s'imposent. Merci donc à eux, merci pour leur accompagnement, leurs discussions, leurs validations et leurs diversions.

En ce qui me concerne, je voudrais remercier en particulier et en premier lieu Stéphane Natkin, mon directeur, qui, au-delà d'une thèse, m'a offert l'opportunité d'un changement de vie en Bretagne. Merci pour sa confiance et pour son suivi, malgré ses engagements professionnels innombrables. Ses questionnements sur le sens de cette thèse au caractère multidisciplinaire, avec cette volonté de toujours la maintenir sur les rails de l'interface homme-machine, m'ont permis de parvenir à un recul que je n'aurais pas su avoir autrement.

Merci à Isabelle Viaud-Delmon, ma co-directrice, pour son investissement malgré les kilomètres d'éloignement. Elle a été pour moi la caution de rigueur et de précision de cette thèse, aussi bien dans l'élaboration de mes protocoles expérimentaux que pour la publication de mes résultats. Ses précieux conseils me resteront sans aucun doute utiles dans les années à venir.

Je remercie aussi tout spécialement Laetitia Gros et Rozenn Nicol, mes encadrantes « de proximité » à Orange. Elles m'auront accompagné dans chacun des événements de cette thèse, et bien plus encore. Elle ont constitué un soutien moral au quotidien, permettant que ces trois dernières années ne soit pas simplement une succession de faits scientifiques, mais aussi une aventure humaine. Merci à elles pour leurs remarques toujours à propos et leur savoir faire expérimental. Merci à Laetitia qui, grâce à son expertise, a su faire de la qualité d'expérience et de la perception auditive des sujets d'étude riches et passionnants, et merci à Rozenn d'en avoir fait de même avec le son 3D, en particulier le son binaural. Enfin, Rozenn, merci pour les découvertes culinaires du mercredi midi !

Je remercie Cécile Le Prado qui m'a également encadré, et qui tout autant que Stéphane m'a suivi bien en amont de cette thèse. C'est aussi grâce à elle que tout a été possible. Merci à elle de m'avoir mis à flots, en particulier les six premiers mois à Paris, et pour ses remarques pertinentes tout au long du chemin.

Par ailleurs, cette thèse a été soutenue grâce à l'engagement et au travail de Mathieu Paquier et Pierre Jouvelot, tous les deux rapporteurs, qui ont su dans des délais serrés me faire parvenir de nombreuses remarques sur mon manuscrit, permettant une discussion intéressante lors de la soutenance. Je n'oublie pas également Alexandre Garcia, président du jury, qui avait également participé à la soutenance à mi-parcours. Je les remercie tous les trois chaleureusement.

De nombreuses autres personnes ont participé directement à l'accomplissement de cette thèse tout au long de ces trois ans. Je remercie les membres de Polymorph et de Studio Anatole, qui ont hautement contribué à la réussite de mon expérience principale, en ayant développé, là aussi dans des délais restreints, un jeu vidéo de A à Z. Merci à Marc Émerit pour avoir rendu son moteur de son binaural compatible sur téléphones mobiles, ce qui lui a sans doute demandé plus d'investissement que prévu. Merci aussi à Thierry Moal de m'avoir suppléé dans la passation de sujets lors de ma dernière expérience. Enfin, je remercie Mathieu Beauval et ses collaborateurs de Radio France, Lidwine Hô de France Télévisions et Pascal Rueff pour les discussions en début de thèse, sur le binaural, sur son utilité et son utilisation concrète, qui ont permis d'irriguer ma réflexion pour la suite.

Je me dois également de remercier tous ceux qui m'ont accompagné quotidiennement pendant et en dehors des heures de travail, ces personnes grâce à qui la thèse passait au second plan le temps d'une pause : l'équipe ILJ au Cnam, l'équipe XDLab à Orange, et bien sûr tous mes comparses thésards et stagiaires du midi à Orange. Je leur adresse un remerciement collégial, de peur d'en oublier certains...mes pensées s'arrêtent sur chacun d'entre eux, qu'ils en soient assurés. Une mention spéciale toutefois aux trois personnes qui tour à tour ont partagé le même bureau que moi, et qui ont dû me supporter un peu plus que les autres : Georges Roussel (qui a soutenu une semaine avant moi, pour l'HDR on se synchronisera encore mieux), Richard Salaün et Mathilde Fernandes. Les meilleurs co-bureau dont on puisse rêver !

Une pensée pour la team du pommier, que je ne pouvais pas ne pas mentionner, Faure, Ferreira, Vilboot et Léo, mes copains de toujours, qui ont été là au cours de cette thèse, en trame de fond la plupart du temps, mais bien présents.

Pour finir, une pensée toute particulière va pour Céline, elle qui a fait intégralement partie de cette aventure, et de toutes les autres, passées et à venir. Rien n'aurait été possible sans son soutien de tous les instants. Et à vous mes enfants, Louis et Anna, arrivés en cours de route, rien ne m'aura procuré plus de joie pendant ces trois ans que d'arriver au travail en retard, encore ensommeillé après une nuit difficile. Il est des bonheurs parfois difficiles à expliquer.

Avant-propos

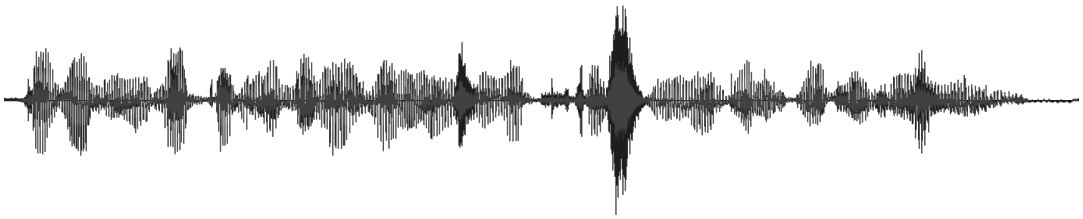
Le titre est le tout premier contact que l'on a avec le sujet d'une thèse. Il concentre en quelques mots l'étendue de la problématique, et c'est à partir de lui que se construisent les premières réflexions. Ainsi, il n'est pas inutile de s'y attarder quelque peu. En guise d'avant-propos, et pour aider le lecteur à mieux cerner le sujet, nous proposons un petit jeu consistant à décortiquer le titre de cette thèse au travers de quelques variations oulipiennes :

0. *Original : Évaluer l'apport du binaural dans une application mobile audiovisuelle.*
1. *Lipogramme en E : Savoir l'apport du binaural sur un Android pourvu d'applications aux stimuli bimodaux.*
2. *Lipogramme en A : Circonscrire l'intérêt d'un son en trois dimensions sur un logiciel mobile pour les oreilles et les yeux.*
3. *Lipogramme en I : Évaluer l'apport d'un son à deux espaces plus un sur un software de smartphone à entendre et à regarder.*
4. *Lipogramme en O : Évaluer l'intérêt du binaural sur un jeu de cellulaire auditif et visuel.*
5. *Lipogramme en A, I, O, U : Déceler l'excellence de l'entente (celle désenchevêtrée en dextre et en senestre, les percepts semblent excentrés de tête) en présence de systèmes de télé-hèlement (genres de télex en externe). Ces systèmes permettent d'entendre, certes, et de présenter en même temps jpeg, etc.*
6. *Définitionnel : Estimer quant à son prix, à sa valeur, à sa quantité, à sa durée la contribution de la technologie qui concerne l'audition par les deux oreilles dans un programme informatique utilisant des fonctionnalités d'une certaine plateforme, comme un système d'exploitation, et opérant sur celle-ci, conçu pour les appareils mobiles, qui regroupe le son et l'image.*
7. *Translation +8 : Évertuer l'apprenti de la binouze dans un appontage modéré augustinien.*
8. *Palindrome strict : Elle. Usi : void ua, Eli bom ? No, it, à cilp, Pa en us n'ad là. Rua ni bu d, trop pal. Reu lavé !*
9. *Négation : Ne pas évaluer l'apport du binaural dans une application mobile.*
10. *Ablation : L'apport du binaural dans une application mobile audiovisuelle.*
11. *Double ablation : L'apport du binaural dans une application mobile.*
12. *Triple ablation : Le binaural dans une application mobile.*
13. *Ajout malheureux : Dévaluer l'apport du binaural dans une application mobile.*
14. *Antonymie : Ignorer l'inutilité de la monophonie en dehors de plusieurs logiciels PC statiques sans son ni image.*
15. *Trop explicite : Évaluer, c'est-à-dire mettre au point une méthode de mesure aboutissant à une note quantitative ou un avis qualitatif (via des données récoltées de façons diverses auprès de sujets), évaluer, disais-je, l'apport, ou l'intérêt,*

ou même plus précisément la qualité d'expérience, qu'on peut envisager globalement ou via des attributs spécifiques (par exemple l'immersion, mais n'anticipons pas), du binaural, ce type de rendu sonore spatialisé, dit « son 3D », qui s'écoute par les deux oreilles (ni plus ni moins) sur un casque, dans une application (ici en l'occurrence un jeu vidéo de type Infinite Runner, qui présente un intérêt du point de vue de son utilisation répandue auprès du grand public et des possibilités de spatialisation sonore) mobile, autrement dit une application installée sur un téléphone mobile, impliquant de prendre en ligne de compte la question des contextes d'utilisation et de leur influence sur l'apport susmentionné, et audiovisuelle, c'est-à-dire, mais dites moi si j'en fais trop, une application qui s'entend et qui se voit, impliquant d'étudier la perception spatiale conjointe des deux modalités, étude pour laquelle nous utiliserons plus à propos l'adjectif auditivo-visuel, voire visuo-auditif.

16. *Réorganisation alphabétique : aaaaaaa bb c ddd eeeeee iiiiii lllllll m nnnn oooo pppp rrr ss tt uuuuuu vv*
17. *Anagramme : Abracadabra ! Ma nouvelle oreille dupa la vieille : nu poupon dû puis instit.*
18. *Défaitiste : Tenter de dire un mot sur l'intérêt tout relatif du son binaural (encore faut-il qu'il soit perçu) dans une application, mais seulement sur Android, et ne sollicitant pas d'autre modalité que la vue et l'audition...*
19. *Ambitieux : Connaître tout de la perception et de la qualité d'expérience sur la totalité des systèmes de restitution sonore dans l'ensemble des terminaux audiovisuels, olfactifs, gustatifs, haptiques et proprioceptifs.*
20. *Jargonneux : Mesuration du contribuement sonoritale du duoaural dans un softiciel mobiliaire à caractère auriculo-œillatoire.*
21. *Alexandrin : L'apport du binaural / dans une application*
22. *Alexandrins ×2 :*
    Évaluer l'apport / du rendu binaural
    Dans une application / mobile et bimodale
23. *Boule de neige :*
    d'
    un
    son,
    dont
    duale
    écoute
    aboutit
    binaural,
    voudrions
    quantifier
    potentielle
    contribution
    audiovisuelle,
    multi-attributs,
    non-immobilement,
    contextuellement,
    expérimentalement.

24. *Fine déduction : L'apport du binaural est une affaire de perception et de contexte d'utilisation.*
25. *Autre point de vue : Et ton manuscrit ça avance ?*
26. *Hétérosyntaxisme : Que le binaural améliore l'expérience mobile audiovisuelle requiert une étude.*
27. *Isoconsonnantisme : Enviant le léopard, de bonne heure allant, deux ânes pleins cassaient ma belle Honda en visant l'aile.*
28. *Isovocalisme : Récap du met : d'abord, du riz au lard, en cubes larges (six rations), rôti. Broccio minute. Prêt !*
29. *Isophonisme : Eve a lu et lappe, hors débine orale, dans Sue ; n'a pli qu'à scier mots. Bile au dit ! Ô visu, hais-le !*
30. *Quadruple contresens : Évaluer la participation financière des deux oreilles dans une assiduité gigotante audiovisuelle.*
31. *Sonore :*



32. *Interrogation : Le binaural apporte-t-il à l'expérience mobile audiovisuelle ?*

Nous laisserons ici le lecteur en proie à cette douloureuse question.

Table des matières

Chapitre 1

Introduction et problématique

1.1 Introduction

1.1.1 Le terminal mobile et le rendu sonore

Les terminaux mobiles représentent aujourd'hui un marché mondial de grande ampleur. Des centaines de millions d'unités vendues chaque année[1], avec une quantité de marques et de modèles toujours grandissante[2]. Ce gigantisme et cette diversité de produits se traduisent techniquement par de nombreuses variations matérielles, permettant à tout utilisateur de choisir son téléphone en passant par tous les stades de la personnalisation. Pour autant, et malgré la diversification conjointe des contenus audiovisuels (films, musiques, jeux vidéo, etc.), l'évolution du rendu sonore se fait plus discret. Encore aujourd'hui, la plupart des smartphones sont monophoniques (ils n'ont qu'une enceinte), ou plus rarement stéréophoniques, pour un rendu relativement peu apprécié (voir par exemple [McMullin et al., 2018]). Certains cas d'utilisation permettent alors d'envisager un raccordement du téléphone à un système d'enceintes plus élaboré, par exemple aux haut-parleurs d'un habitacle de voiture, ou à un réseau d'enceintes bluetooth. Néanmoins, la pluralité des contextes rencontrés par un utilisateur de smartphone ne permet pas toujours cette extension : milieu en extérieur, avec du bruit ambiant, en présence de personnes susceptibles de la percevoir comme une nuisance sonore, ou de capter des informations personnelles, contexte impliquant un déplacement physique de l'utilisateur, etc. Cette question des contextes, variés, dynamiques et imprévisibles, constitue dès lors une pierre d'achoppement à une écoute de qualité.

Dans ces conditions, l'utilisation d'un casque (ou d'écouteurs) semble être une solution pour remédier à ces problèmes : non seulement elle isole l'utilisateur du contexte extérieur (et inversement, elle soulage l'entourage des sons indésirables émis par le téléphone), mais elle est aussi une façon d'accéder à un rendu de meilleure qualité, selon le casque, et indépendamment du téléphone. Si l'écoute sur casque induit intuitivement un rendu au mieux stéréophonique (1 canal pour chaque oreille), des solutions

1. https://www.zdnet.fr/actualites/chiffres-cles-les-ventes-de-mobiles-et-de-smartphones-39789928.htm
2. 86 marques recensées par https://www.gsm55.com/liste-marques-mobile, et plusieurs centaines de modèles

d'écoute spatialisée adaptées à ce support ont vu le jour depuis quelques décennies. Le rendu en binaural en fait partie, qui permet de spatialiser un son n'importe où à l'extérieur de la tête de l'utilisateur, tout autour de lui.

1.1.2 Le binaural peut-il améliorer l'expérience audiovisuelle ?

Le binaural est une technologie de rendu qui se déploie sur deux canaux, un pour l'oreille droite, un pour l'oreille gauche. Il s'appuie sur l'écoute naturelle de l'être humain pour imprimer au son les indices de localisations nous permettant d'identifier sa provenance spatiale [Blauert, 1997]. Grâce à ce principe, le binaural peut spatialiser un son n'importe où autour de l'auditeur avec un simple casque. Pas besoin donc de matériel coûteux dans sa mise en œuvre, argument qui plaide de surcroît en faveur d'une diffusion grand public.

Mais indépendamment de ces qualités sonores propres se pose aussi la question de l'association avec un visuel lors d'une utilisation sur smartphone. Malgré la qualité des écrans toujours plus grande, notamment leur résolution, la nécessité de les transporter partout contraint leur taille à une certaine limite (les plus grands écrans ne dépassent pas les 16cm de diagonale). Le binaural offre l'opportunité de déployer la scène au delà de cet écran, non seulement en spatialisant des éléments hors-champ, mais aussi en amplifiant par le son les éléments présents à l'écran. Intuitivement, cette association pourrait aussi bien enrichir la qualité d'expérience de l'utilisateur (notamment en termes d'immersion) que la dégrader (susciter un certain inconfort, par exemple par la dissociation entre son et image, à l'image du *virtual motion sickness* en contexte de réalité virtuelle, provoqué par la perception d'un mouvement du la scène visuelle alors que la personne est immobile, ou par un retard entre ses propres mouvements et ceux du visuel [Hettinger et Riccio, 1992]).

Au delà de l'intuition, l'objectif principal de ce travail est d'identifier les apports potentiellement multiples du binaural, puis de mesurer leurs conditions d'apparition. Néanmoins, le fait que les terminaux mobiles s'utilisent dans de multiples contextes nécessite de se pencher sur la méthode expérimentale. Une simple expérience en laboratoire prendrait le risque de ne pas représenter suffisamment la diversité de ces contextes et leur influence sur la qualité d'expérience. L'élaboration et la mise en œuvre du protocole constituent donc également une étape importante du travail de recherche.

Enfin, avant d'entrer dans le vif du sujet, une étape préliminaire du travail sera d'étudier le lien perceptif qui unit un son binaural à un visuel sur mobile. En effet, pour mesurer l'apport du binaural, encore faut-il s'assurer de proposer une expérience dans laquelle le son et image sont correctement associés.

1.1.3 La perception spatiale d'un son binaural couplé à un visuel sur terminal mobile

Dans la vie quotidienne, nous sommes peu habitués à visualiser des contenus sur un petit écran et entendre en même temps des sons avec un tel niveau de spatialisation. La question qui d'emblée peut frapper le lecteur est celle de la coïncidence spatiale entre le son et l'image. Où se trouve la bonne position pour un son dont le correspondant visuel ne dépasse pas les quelques centimètres ? Imaginons par exemple un visuel spectaculaire, une explosion, ou un véhicule allant à toute vitesse, faut-il placer la source sonore en face de l'utilisateur à 30cm, à l'endroit même du visuel ? Ou décorréler complètement les deux modalités, et le plonger au cœur d'une scène sonore, disproportionnée par rapport au visuel ? Notons bien que la question n'est pas de savoir si une solution est artistiquement meilleure qu'une autre, car tout type d'agencement spatial pourrait trouver sa justification selon l'application. Il s'agit plutôt de savoir si, pour un rapport de position donné, son et image sont correctement associés d'un point de vue perceptif par l'utilisateur.

Quoi qu'il en soit, avant même de mesurer l'apport potentiel du binaural, il convient donc de répondre à cette question, au moins partiellement, afin de savoir le niveau d'inconfort perceptif qu'engendre une présentation donnée.

1.2 Problématique

L'objectif général de cette thèse est de déterminer l'apport du binaural dans une application mobile audiovisuelle. Néanmoins, nous voyons avec ce qui précède qu'il ne peut pas être atteint sans passer par des étapes intermédiaires de réflexion. Plusieurs problèmes ont été soulevés, qui méritent d'être traités indépendamment et de façon successive. Si l'on reprend simplement ce sujet « Évaluer l'apport du binaural dans une application mobile audiovisuelle », on recense les six mots à partir desquels la problématique va finalement se construire. Une partie importante de ce travail de thèse a consisté à articuler des axes de réflexion à partir de ces mots, et à en dégager une problématique cohérente et ordonnée. Nous épargnons au lecteur ici le cheminement quelque peu chaotique qui y a mené, et présentons de façon succincte ces mots remis dans l'ordre logique qui présidera à la suite du texte : il y a tout d'abord « binaural », la technologie à laquelle on s'intéresse, et « audiovisuelle », qui insiste sur le couplage bimodal entre son et image, et la composante perceptive qui s'y rattache. Vient ensuite « apport », qui évoque la notion générale de qualité d'expérience et rattache la perception du contenu audiovisuel à un ressenti utilisateur. « Évaluer », le verbe, introduit la méthode pour caractériser cette qualité d'expérience, et définit donc la tâche à accomplir. « Mobile » introduit la dépendance de la qualité d'expérience et de la méthode d'évaluation au contexte. Enfin, le terme « application » est le support concret de notre investigation, qui permet aussi de délimiter son champ d'action.

Le découpage qui suit en cinq sous-parties reprend ce cheminement, les quatre premières associant une question spécifique à chaque mot ou groupe de mots évoqués, et la cinquième reprenant la question principale de la problématique.

1.2.1 Question préliminaire : quelle perception spatiale a-t-on d'un son binaural lorsqu'il est couplé à une image ?

La conception d'une application audiovisuelle implique une mise en scène des sources sonores et visuels qui la composent. Son premier objectif, avant toute visée artistique, est que l'utilisateur perçoive les objets audiovisuels selon la volonté initiale du concepteur, en associant avec succès chaque son et chaque image qui vont ensemble. Pour que cette mise en scène soit perceptivement cohérente, il faut donc se pencher sur la relation spatiale entre les deux modalités et comprendre un peu mieux la façon dont l'être humain les perçoit conjointement. Néanmoins, le sujet est vaste et pourrait à lui seul remplir plusieurs manuscrits de thèse. Il nous faut nous concentrer sur ceux de ces aspects perceptifs qui peuvent avoir une influence sur le couplage son binaural-visuel sur mobile. Notons bien ici que nous nous intéressons exclusivement au rapport qu'entretiennent le son et l'image d'un point de vue spatial. D'autres études seraient possibles (notamment sur leur relation sémantique). Cette limitation est motivée par le fait que le binaural nous intéresse ici en sa qualité de rendu sonore spatialisé.

Par ailleurs, le rendu binaural a pour particularité d'être individualisé, c'est-à-dire adapté à un seul et unique utilisateur, en se basant sur les indices de localisation déterminés par sa morphologie. Pour les autres, le rendu peut être altéré et provoquer des défauts de localisation dans certaines directions. Idéalement, il faudrait donc que tous les contenus binauraux produits, qu'ils soient audiovisuels ou non, soient personnalisables pour que chaque utilisateur vienne y greffer quand il veut ses propres indices de localisation. Or le processus d'acquisition de ces indices, comme nous allons le voir par la suite, est coûteux et inaccessible au grand public. Pour répondre pleinement à la question de l'interaction son-image dans la perception spatiale, il faut donc également s'intéresser à l'influence de l'individualisation.

Enfin, concernant la perception spatiale, une autre piste de réflexion possible est tracée par les pratiques les plus usitées actuellement. Des contenus audiovisuels avec binaural sont déjà produits depuis plusieurs années. Les techniciens en charge de ces contenus (ingénieurs du son, monteurs, etc.), sans pour autant avoir théorisé leurs pratiques, ont adopté des solutions qu'il sera intéressant d'analyser à la lumière de nos propres réponses expérimentales.

1.2.2 Quels types d'apports possibles pour le binaural ?

Le binaural ajouté à un contenu audiovisuel sur mobile est susceptible d'enrichir l'expérience de l'utilisateur. Voilà l'hypothèse sur laquelle se fonde ce travail de thèse. Mais de quelle façon l'enrichit-elle ? Les attributs visés peuvent être d'ordre psychologique (e.g., l'immersion) ou physiologique (e.g., les réflexes). Pour les identifier précisément, on s'appuiera sur les travaux déjà menés sur le binaural, et plus généralement dans le domaine du son spatialisé. Les apports identifiés conditionneront le protocole expérimental et les supports audiovisuels utilisés (la ou les applications), mais seront aussi conditionnés par l'équipement à disposition (par exemple, difficile de tester la capacité d'un sujet à naviguer dans un espace grâce au son binaural si on ne dispose pas d'une salle prévue à cet effet).

1.2.3 Quelle méthodologie pour mesurer ces apports dans le contexte d'utilisation d'un terminal mobile ?

Les contextes multiples d'utilisation des applications mobiles requièrent de se pencher sur la question méthodologique. La plupart des expériences d'évaluation de qualité d'expérience se passent en laboratoire, milieu peu représentatif des contextes d'utilisation d'un mobile, de leur diversité et de leur dynamisme. Nous souhaitons à l'inverse que notre cadre expérimental se rapproche d'une utilisation réelle. Dans un premier temps, il faut définir la notion de contexte dans son acception la plus générale, puis concentrer le champ de recherche sur les contextes mobiles en particulier. La notion de mobilité (i.e., le déplacement de l'utilisateur), spécifique à l'utilisation d'un smartphone, doit retenir l'attention.

Par la suite, le protocole expérimental doit tenir compte de ces aspects contextuels potentiellement influant sur la perception du binaural. L'élaboration de ce protocole, la réflexion dont il procède, s'ils n'apportent pas directement d'information sur l'apport du binaural, n'en restent pas moins de première importance dans le cadre de ce travail. Ils sont une façon d'approfondir la facette méthodologique de la problématique. De ce point de vue, l'inclusion du contexte à l'expérience peut se révéler d'un intérêt double : elle permet d'obtenir des informations a priori plus représentatives d'une utilisation réelle du smartphone, mais aussi de mesurer la relation entre contexte et qualité d'expérience. La confrontation entre les méthodes traditionnelles en laboratoire et les solutions proposées ici, non seulement d'un point de vue théorique, mais également dans leurs aspects les plus concrets, a pour but d'apporter quelques lumières aux futurs expérimentateurs du domaine de la qualité d'expérience sur mobile.

1.2.4 La mise en œuvre : sur quelle application ?

Une fois les apports potentiels du binaural identifiés et la méthode expérimentale choisie, il s'agit de déployer l'expérience. Ici se pose une dernière question d'importance, celle du choix de l'application sur laquelle mener les expériences. Une question à la croisée des chemins entre la réflexion sur les applications utilisées par le grand public, leurs potentiels enrichissements par le binaural, leurs rapports au contexte, et d'autre part des considérations pratiques guidées par les contraintes de temps imparti par la thèse.

1.2.5 Question finale : quel est l'apport du binaural dans une application mobile audiovisuelle ?

La réponse à cette question finale constitue bien évidemment la ligne d'arrivée. Notons d'ores et déjà sur ce point qu'avec tout ce qui a été évoqué, les différents types d'apports possibles du binaural, les influences potentielles diverses et variées du contexte, l'interaction perceptive entre son et image, le nombre faramineux d'applications différentes sur lesquelles déployer l'expérience, il faudra nécessairement se restreindre non seulement sur les choix expérimentaux, mais aussi probablement dans l'analyse des données. Ce travail constitue donc avant tout un défrichage de ce qui nous aura paru

le plus pertinent dans un premier temps, et fournit des pistes de réflexions pour des approfondissements futurs.

1.3 Plan du manuscrit : deux parties, l'état de l'art divisé en deux

Nous présentons ici le plan du manuscrit qui découle des cinq questions de la problématique. Nous proposons un découpage en deux parties, afin de séparer les aspects de perception à proprement parler de l'évaluation du binaural sur mobile. Dans la première partie, le chapitre 2 propose un premier état de l'art consacré d'une part au rendu binaural, sa définition, son enregistrement et sa reproduction, la notion d'individualisation, et d'autre part à la perception auditivo-visuelle. Il s'agira principalement de se concentrer sur la perception spatiale. Le chapitre 3 présente une expérience visant à mieux comprendre la perception spatiale d'un son binaural couplé à un visuel sur mobile. Le choix de la méthode y est discuté, puis l'expérience est présentée, son protocole, son déroulement, les résultats obtenus et l'analyse qui en découle.

Dans la seconde partie, le chapitre 4 développe un nouvel état de l'art portant sur la qualité sonore et la qualité d'expérience, leurs définitions et leurs méthodes d'évaluation. La notion de contexte est ensuite introduite comme élément à prendre en compte dans l'évaluation de la qualité d'expérience, en particulier via la méthode de déploiement. Le chapitre 5 est une transition permettant de synthétiser et justifier les éléments retenus dans l'état de l'art pour l'expérience à suivre. Il est l'occasion aussi d'aborder le choix de l'application, dont la scène virtuelle, avec notamment la relation spatiale entre objets sonores et visuels, sera construite à la lumière des conclusions tirées au chapitre 3. Le chapitre 6 est une présentation de l'expérience, son protocole, son déroulement, les résultats et l'analyse qui en découle. Les données aussi bien que la méthodologie sont discutées. Enfin, la Conclusion, chapitre 7, est l'occasion de revenir sur la problématique générale de la thèse, tenter d'y répondre et de proposer des pistes de réflexion pour de futures recherches.

Première partie

Perception auditivo-visuelle

Chapitre 2

Binaural et perception auditivo-visuelle

2.1 Le binaural, définition

Les systèmes de reproduction sonore spatialisée sont des technologies permettant de simuler à l'écoute humaine la présence de sources sonores virtuelles localisées dans l'espace. Quand un son monophonique est émis par une enceinte, la position de la source sonore identifiée par l'auditeur est l'enceinte elle-même. Avec les techniques de rendu spatialisé, à commencer par la stéréo [Blumlein, 1958], il est possible de simuler un emplacement différent du système lui-même, et de créer ainsi des sources virtuelles. Si la stéréo permet de créer une source sonore positionnée sur un plan horizontal entre deux enceintes émettrices, d'autres techniques plus récentes permettent d'élargir le champ sonore, comme par exemple le rendu 5.1 avec cinq enceintes, une frontale, deux latérales avant et deux latérales arrière, plus un caisson de basses (une enceinte dédiée aux basses fréquences), permettant de positionner une source tout autour de l'utilisateur sur un plan horizontal [ITU-R, 2012]. En multipliant ainsi les enceintes autour de l'utilisateur, au-dessus, en dessous, etc., on multiplie les positions possibles pour une source sonore donnée, jusqu'à atteindre la sphère d'enceintes entourant complètement l'utilisateur (comme par exemple le système de restitution Ambisonic Figure 2.1, développé par le laboratoire d'acoustique du CNAM, à Paris, avec 50 enceintes [Lecomte et al., 2015]).

Plus le système est sophistiqué, plus il peut d'une part favoriser l'efficacité à localiser les sources, et d'autre part potentiellement renforcer la crédibilité ou le réalisme de la scène sonore, sa lisibilité, et amener à une sensation d'immersion plus intense de la part de l'auditeur. Dans [Pulkki et Hirvonen, 2005] par exemple, les auteurs comparent la capacité à localiser avec des systèmes à 5 ou 8 canaux. Ils montrent que le système à 5 canaux (identique à un système 5.1 privé de son caisson de basses), à l'inverse du système à 8 canaux, est incapable de simuler des sources virtuelles dans certaines directions (au delà de 70 ° du plan médian, avec les systèmes de reproduction spécifiques testés par les auteurs). [Hamasaki, Nishiguchi et al., 2004] et [Hamasaki, Y. Nakayama et al., 2007] présentent tous les deux un système de reproduction sonore en 22.2 et le comparent avec des systèmes stéréo et 5.1, le premier en termes de sentiment de présence, et le second selon divers attributs spatiaux (« profondeur », « caractère dynamique », « ampleur », « naturel », etc.). Les résultats indiquent de

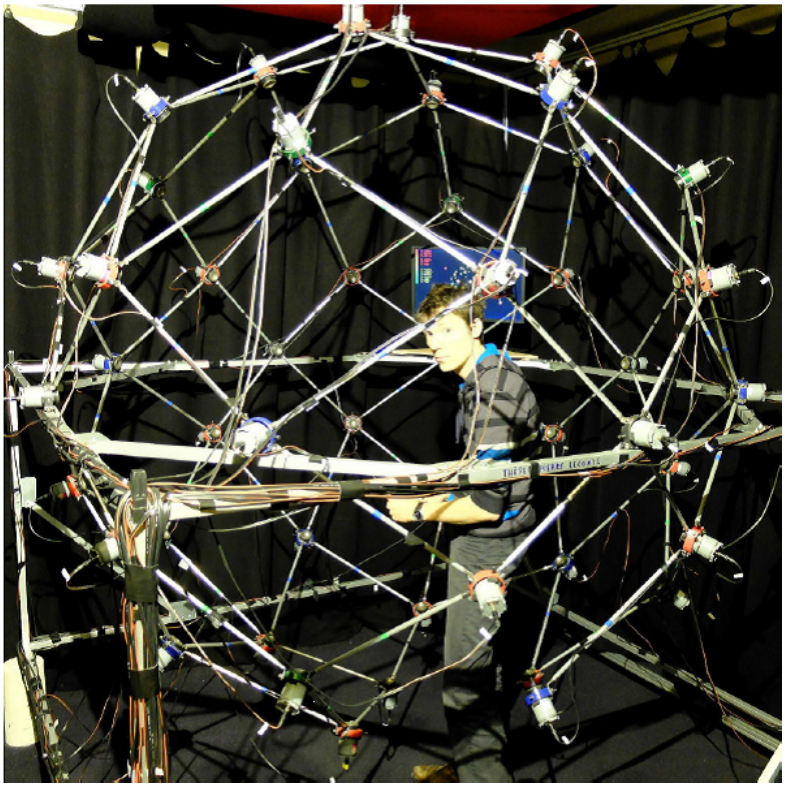


FIGURE 2.1 – Une installation sonore au Cnam de Paris avec 50 enceintes distribuées sur une sphère pour la restitution Ambisonic

meilleures performances du système 22.2 pour les deux publications. Cependant, si certains systèmes ont su se faire une place dans les foyers, la plupart d'entre eux ont l'inconvénient de nécessiter un matériel cher, pour une mise en place encombrante, et sont par conséquent inaccessibles au grand public.

Sur smartphone, le problème est d'autant plus marqué à cause de son caractère portatif. Les appareils ne possèdent en général qu'une seule enceinte (quelques modèles comportent deux enceintes, mais très rapprochées l'une de l'autre), privant de la possibilité de spatialiser un son, même a minima. La seule solution serait un raccordement à un groupe d'enceintes externes, opération rendue fastidieuse par la multiplicité des contextes d'utilisation et leurs changements fréquents, en particulier lorsqu'ils impliquent un mouvement de l'utilisateur[1]. On privilégie alors plutôt le casque audio, accessoire adéquat dans tous les contextes, en intérieur ou extérieur, en transport, en déplacement physique ou immobile, etc. Non seulement il permet de restreindre la zone d'écoute à l'utilisateur (évitant de déranger les personnes alentour), mais il permet en plus le déploiement d'une technologie de son spatialisé qui lui est propre : le rendu binaural.

Le binaural est la seule technique de restitution sonore spatialisée qui ne nécessite pas de matériel supplémentaire. Avec un casque, un signal monophonique ou stéréophonique placerait la source sonore dans la tête de l'auditeur. Le binaural permet d'externaliser cette source n'importe où autour de lui [BLAUERT, 2013]. Il reproduit

1. Pour autant il y a des solutions aujourd'hui qui se développent, enceintes connectées, multiroom, sound zones, etc., plutôt adaptées à une utilisation sédentaire du téléphone

la transformation qu'un signal subit dans l'interaction de l'onde acoustique avec la morphologie de l'auditeur. Plus précisément, c'est la distance qui sépare les deux oreilles, la forme des pavillons auriculaires et l'écran acoustique constitué par la tête et le torse qui sont à l'origine de cette transformation. Techniquement, le binaural peut être considéré comme un enrichissement de la stéréo en ce qu'il ne nécessite que deux signaux, un pour chaque oreille, et qu'il intègre des indices de localisation similaires : la différence de temps interaurale (ou ITD pour *Interaural Time Difference*) et la différence de niveau interaurale (ou ILD pour *Interaural Level Difference*). Ces deux indices traduisent le fait qu'une onde sonore émise par une source placée à un endroit donné de l'espace n'arrive pas en même temps aux deux oreilles, ni au même niveau sonore. Ils facilitent principalement la localisation en azimut (i.e., différencier la gauche de la droite, voir à ce propos [Wightman et Kistler, 1992]). Cependant, à la différence de la stéréo, le binaural cherche à reproduire l'ITD et l'ILD « naturelles » d'un auditeur, alors que la stéréo s'appuie sur des différences de temps et d'amplitude non-personnalisées. Par ailleurs, le binaural intègre en plus un autre indice de localisation, le filtrage spectral que subit le son en interagissant avec le pavillon d'oreille, organe fait de cavités et de bosses qui déforment le son différemment selon la provenance de la source ([Spagnol, Geronazzo et Avanzini, 2013], voir aussi [Algazi, Duda, Duraiswami et al., 2002] sur le rôle du torse et de la tête). Ce filtrage, différent pour l'oreille droite et l'oreille gauche, favorise principalement la localisation en élévation (au-dessus, en dessous) et en profondeur (devant, derrière) [Asano, Yoiti Suzuki et Sone, 1990 ; Langendijk et Bronkhorst, 2002 ; Zhang et Hartmann, 2010]). On le définit comme une fonction de transfert du son allant d'un point d'émission donné jusqu'à l'entrée du conduit auditif de l'auditeur, qu'on appelle couramment HRTF, pour *Head-Related Transfer Function*. Le binaural consiste donc à reproduire une scène auditive de la façon la plus réaliste possible spatialement, en filtrant chaque source sonore par le couple de HRTF qui correspond à sa position.

Il existe deux façons de créer un son binaural. La première consiste à enregistrer naturellement une scène sonore en binaural, et à la restituer telle quelle au casque. L'acquisition se fait en insérant deux microphones de petite taille dans le creux des oreilles qui capturent le son déjà enrichi de tous ses indices de localisation. On peut aussi enregistrer le son à l'aide d'une tête artificielle dans laquelle est embarqué un système d'enregistrement plus sophistiqué (voir la Figure 2.2).

L'autre solution consiste à créer un effet binaural de toute pièce à partir d'un son monophonique, en lui appliquant artificiellement les filtres HRTF correspondant à une direction de l'espace. Cette technique s'appelle la synthèse binaurale. L'avantage est qu'elle permet de construire une scène sonore spatialisée fictive, par exemple une scène de jeu vidéo ou de film, quand le binaural naturel ne peut capturer que des scènes sonores réelles.

Les performances du binaural en termes de spatialisation ont été examinées à plusieurs reprises. Dans son état de l'art, [Bahu, 2016, p. 27] résume les tests menés qui comparent la localisation de sources rendues en binaural avec celle de sources réelles en champ libre (c'est-à-dire dont les ondes émises ne subissent aucun effet de salle, pas de réverbération ou de diffraction). La majorité des études reportent une précision identique pour les deux [Wightman et Kistler, 1989 ; Møller, Sørensen, Jensen et al., 1996 ; Langendijk et Bronkhorst, 2000 ; R. L. Martin, McAnally et Senova, 2001]. Parmi les quelques défauts toutefois observés, quelques études

FIGURE 2.2 – Un exemple de tête artificielle, la KU 100 de chez Neumann.

rapportent une augmentation des confusions avant-arrière [Wightman et Kistler, 1989 ; Bronkhorst, 1995 ; Kim et Choi, 2005]. Cette confusion est expliquée par le fait que l'écoute naturelle s'accompagne généralement de petits mouvements de la tête, conscients ou non, qui facilitent la localisation, tandis que le binaural n'en tient généralement pas compte [2]. Un autre défaut reporté est une légère dégradation de localisation en élévation [Wightman et Kistler, 1989 ; Bronkhorst, 1995], attribuée à l'existence de distorsion dans les HRTF mesurées. Quoi qu'il en soit, et malgré ces quelques défauts, les résultats témoignent de la qualité de spatialisation du binaural.

Néanmoins, une des particularités du rendu binaural est son caractère individualisé. Non seulement les indices de localisation dépendent de la position de la source sonore, mais ils dépendent aussi de la morphologie de l'auditeur. Une scène sonore binaurale présentée à un auditeur est donc faite sur mesure pour lui. Tout autre écoutant, réceptionnant en quelque sorte les sons avec d'autres oreilles que les siennes, localisera potentiellement moins bien les sources. [Mrsic-Flogel et al., 2001] mettent en évidence cette différence de perception au niveau neuronal en montrant la différence d'activité cérébrale d'un auditeur entre des stimuli binauraux individualisés et non-individualisés. Par ailleurs, sur le plan comportemental, [Møller, Sørensen, Jensen et al., 1996] font par exemple état d'erreurs dans le plan médian et de davantage de confusions avant-arrière. Si ce processus d'individualisation semble nécessaire à une écoute optimale, il pose le problème de la distribution d'une telle technologie auprès du grand public.

Dans le cas du binaural naturel, le problème est difficile car une scène enregistrée par une paire d'oreilles ne peut plus être réajustée a posteriori pour un autre auditeur. Pour contourner ce problème, la tête artificielle est censée incarner une morphologie moyenne pour convenir à une majorité d'auditeurs. Néanmoins, des recherches montrent que cette solution n'égale pas l'enregistrement individualisé en termes de localisation [Møller, Hammershøi et al., 1999 ; Minnaar et al., 2001]. Des conclusions similaires sont formulées dans [Møller, Jensen et al., 1996], qui cherchent à sélectionner un profil « type » permettant aux membres d'un petit groupe d'individus de localiser efficacement les éléments d'une scène sonore. Ces résultats sont probablement dus en partie à une disparité morphologique importante entre les individus [Møller, Sørensen, Hammershøi et al., 1995] qu'un unique représentant ne sera pas apte à couvrir.

Pour la synthèse binaurale en revanche, il est possible d'adapter le jeu de HRTF à son auditeur. Néanmoins, le processus d'acquisition des HRTF est fastidieux et coûteux [Algazi, Duda, Thompson et al., 2001 ; Warusfel, 2003 ; Carpentier et al., 2014 ; Rugeles Ospina, Emerit et Daniel, 2015]. En effet, dans les techniques actuelles de mesure, le matériel requis inclut des microphones à insérer dans les conduits auditifs (voir Figure 2.3, ce qui nécessite parfois de faire en amont un moulage du conduit auditif chez un audioprothésiste), un système d'enceintes permettant d'émettre tout autour de l'utilisateur (pour mesurer les HRTF dans toutes les directions de l'espace, y compris en dessous), un système de *head tracking* pour

2. Il existe des systèmes de *head tracking* qui, fixés à la tête, permettent une synthèse binaurale qui s'adapte en temps réel à ses mouvements. Grâce à ce système, la localisation s'en retrouve grandement améliorée [Faure et PALLONE, 2005], proche même d'un espace réel en termes de navigation [Picinali et al., 2014]. Mais étant donnée la singularité du matériel requis, nous n'envisageons pas son emploi dans cette thèse.

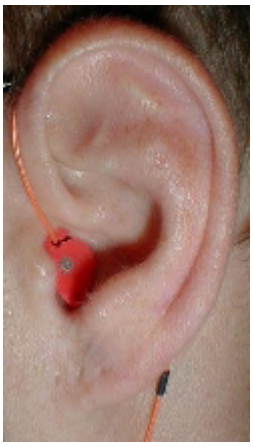


FIGURE 2.3 – Un exemple de microphone inséré dans un conduit auditif. Photo prise lors d'une campagne de mesures de HRTF menée par l'IRCAM.

tenir compte des mouvements intempestifs du sujet pendant la mesure, etc. De plus, les HRTF doivent être indépendantes de tout effet de réverbération de salle, impliquant d'exécuter la manœuvre en salle anéchoïque (voir Figure 2.4). Enfin, le temps de mesure peut être long (allant de 2h à une vingtaine de minutes) selon le matériel, notamment le nombre d'enceintes à disposition, et la méthode utilisée.

Différentes solutions ont été examinées pour pallier ce problème, à commencer par l'utilisation de HRTF standard mesurées sur des têtes artificielles. Sans surprise, on retrouve les mêmes problèmes de localisation que pour l'écoute d'un son binaural naturel enregistré avec une tête artificielle [WENZEL et al., 1993]. L'utilisation d'un jeu préféré de HRTF non-individuelles, commun à un groupe de personnes et sélectionné dans une base de données déjà établie, n'est pas non plus idéale ; [SCHÖNSTEIN et KATZ, 2012] en font l'expérience et montrent une trop forte variabilité inter-sujets (cependant moins prononcée pour des auditeurs experts). Dans la vue d'ensemble récente qu'ils proposent, [GUEZENOC et SEGUIER, 2018] font état de nombreuses autres solutions, séparées en quatre catégories :

— améliorer les conditions de la mesure acoustique : réduction du temps et du coût matériel, précision accrue des mesures en tenant compte notamment des mouvements indésirables du sujet pendant la session, etc. ;

— calculer les HRTF individuelles par une simulation numérique : on modélise la propagation des ondes acoustiques aux abords du sujet représenté en 3D ;

— calculer les HRTF individuelles en faisant mathématiquement le lien entre la morphologie d'un sujet, modélisée grâce à des mesures anthropométriques, et les HRTF ;

— sélectionner un jeu de HRTF non-individuelles dans une base de données en se basant sur des retours perceptifs du sujet, ou calculer les HRTF individuelles en partant d'un jeu de HRTF non-individuelles pré-sélectionné, grâce à une succession de retours perceptifs du sujet.

Néanmoins, l'amélioration des conditions de la mesure acoustique ne permet pas encore de s'affranchir d'une salle anéchoïque et d'un matériel toujours coûteux. La simulation numérique a toujours ce désavantage d'être coûteuse en temps (temps de calcul long) et en matériel (nécessité d'avoir un scanner 3D). De plus cette méthode, malgré

FIGURE 2.4 – Le système d'enceintes utilisées à Orange Labs Lannion pour mesurer les HRTF en salle anéchoïque. Pendant la mesure l'utilisateur, assis sur une table tournante, fait un tour complet sur lui-même, de façon à ce que les demi-cercles d'enceintes puissent émettre tout autour de lui.

des résultats prometteurs, manque encore d'études perceptives pour être validée. Il en va de même pour les deux suivantes, pour lesquelles les performances de localisation sont de surcroît toujours en deçà de celles obtenues avec des HRTF individuelles.

Une autre piste consisterait à simplement se passer d'individualisation. Plusieurs recherches font état de la capacité d'un sujet à apprendre à localiser avec d'autres oreilles que les siennes. Dans [Hofman, Van Riswick et Van Opstal, 1998], les auteurs changent physiquement la forme des oreilles de sujets, en faisant porter aux sujets des moules en cire sur leurs pavillons auriculaires. Cette modification provoque dans un premier temps des difficultés à localiser les sons, mais après une période d'accoutumance, ceux-ci réapprennent à localiser aussi efficacement qu'avec leurs oreilles normales. Chose intéressante, après cet apprentissage les sujets conservent la capacité à localiser avec leurs oreilles normales une fois les moules ôtés, sans nécessité de ré-entraînement. De façon analogue, cet apprentissage semble possible avec des jeux de HRTF standard [Balan et al., 2015]. L'entrainement se fonde alors sur une tâche de localisation immédiatement suivi d'une indication visuelle [Zahorik et al., 2001 ; Shinn-Cunningham, Streeter et Gyss, 2005] ou proprioceptive [Parseihian et Katz, 2012 ; Honda, Shibata, Hidaka et al., 2013] sur l'erreur commise. Dans [Honda, Shibata, Gyoba et al., 2007], les auteurs testent les capacités de localisation de sujets après avoir joué à un jeu vidéo, et obtiennent des résultats de même teneur pour tous les sujets, avec ou sans HRTF individuelles. Ces études mettent en avant les interactions multisensorielles qui régissent notre perception spatiale au quotidien. Elles suggèrent que dans une tâche multisensorielle incluant du son binaural, le repérage spatial peut se faire sans indices de localisation individualisés, à partir du moment où un apprentissage multisensoriel et reposant sur l'action est autorisé. Dans l'étude récente [Nicol, Emerit et Gros, 2018], les auteurs évaluent la performance de localisation de sujets selon différents jeux de HRTF –incluant les leurs– dans un contexte de réalité virtuelle. Les résultats indiquent que moins d'un tiers des sujets localisent mieux avec leurs propres HRTF.

Pour ce qui concerne directement le sujet de cette thèse, dans le contexte d'une application audiovisuelle, quel que soit le type de binaural utilisé, il faut tenir compte de l'interaction des modalités sensorielles entre elles. Notre volonté étant bien de mesurer l'apport du binaural tel qu'il se présenterait dans une utilisation quotidienne d'un smartphone, la question de l'individualisation du binaural, des difficultés techniques qu'elle pose et de la possibilité de s'en affranchir est donc importante. Dans la section qui suit, nous nous penchons plus en détail sur la relation multisensorielle entretenue entre vision et audition.

2.2 La perception auditivo-visuelle

Bien qu'il existe des exemples d'applications purement auditives, la majorité des applications possèdent un visuel, que celui-ci soit animé ou non, en relation sémantique avec l'audio ou non. La modification de la modalité sonore par ajout du binaural peut avoir des répercussions sur la perception de ce visuel. L'objectif de cette section est de mettre en lumière les interactions entre modalités sensorielles visuelle et auditive, en particulier celles qui peuvent apparaître avec l'ajout du binaural. Nous présentons en premier lieu ces interactions telles qu'elles se produisent au niveau neuronal, pour

« remonter » jusqu'au niveau comportemental, afin de nous focaliser sur l'effet qui nous paraît déterminant pour notre sujet, l'effet ventriloque. Introduisons en premier lieu la notion d'intégration sensorielle, notion cruciale pour décrire notre perception unifiée, mais néanmoins multisensorielle, du monde.

2.2.1 Du point de vue neuronal au comportemental, présentation succincte

L'intégration sensorielle est cette capacité que nous avons de percevoir le monde de façon unifiée, bien qu'il nous parvienne par de multiples informations sensorielles (visuelles, auditives, proprioceptives, olfactives, gustatives, vestibulaires, etc.) Dans [Ernst et Bülthoff, 2004], les auteurs nous disent que cette intégration décrit au niveau perceptif les interactions entre les signaux redondants du monde extérieur, c'est-à-dire par exemple les signaux qui sollicitent notre perception auditive, visuelle ou haptique et qui proviennent d'un même objet ou d'un même événement extérieur.

Si ces interactions font état d'une collaboration entre les différentes modalités sensorielles à un niveau supérieur de l'activité cérébrale, il a été découvert depuis plusieurs décennies que des mécanismes d'intégration multisensorielle existent aussi dès le niveau neuronal. [Meredith et Stein, 1983 ; King et Palmer, 1985] montrent que certains neurones chez les mammifères, situés dans une région du cerveau appelée colliculus supérieur, et aptes à répondre aussi bien à des stimuli auditifs que visuels, ont une activité différente quand les deux modalités leur sont présentées en même temps. Ce phénomène est appelé « renforcement » ou « affaiblissement multisensoriel », selon que la réponse de la cellule nerveuse s'accroît ou s'atténue avec l'ajout d'une modalité (nous empruntons le terme français de renforcement multisensoriel à [Gueguen, 2011], pour traduire de l'anglais « crossmodal enhancement »). En particulier, il a été montré que quand un des stimuli, ou même les deux stimuli, ne sont pas assez saillants pour assurer leur détection de manière unimodale (image floutée, ou signal sonore bruité par exemple), le renforcement est alors accru et facilite la détection (voir Figure 2.5). Si ce renforcement a lieu en particulier quand les stimuli sont localisés au même endroit de l'espace, l'affaiblissement apparaît quant à lui quand ils proviennent de positions différentes. Dans [Bell et al., 2005], les auteurs font également état d'une vitesse de réponse neuronale accrue dans le cas des stimuli audiovisuels spatialement alignés, par rapport au visuel seul. Tous ces résultats indiquent que les neurones multisensoriels participent à notre représentation de l'espace [Alais, Newell et Mamassian, 2010].

Par la suite, ces neurones multisensoriels ont été identifiés dans d'autres zones du cerveau [Alais, Newell et Mamassian, 2010]. Dans le cortex par exemple, l'intégration multisensorielle est favorisée lorsque les signaux proviennent de stimuli familiers et sémantiquement cohérents (un visuel de chien avec un son d'aboiement plutôt qu'un son de miaulement par exemple). On associerait alors ici ce renforcement au décodage sémantique des objets perçus, plutôt qu'au repérage spatial [Barraclough et al., 2005].

Cette réponse neuronale multisensorielle, qu'elle soit renforcement ou affaiblissement, est elle-même sujette à des variations. Selon [Meredith et Stein, 1983], la raison de

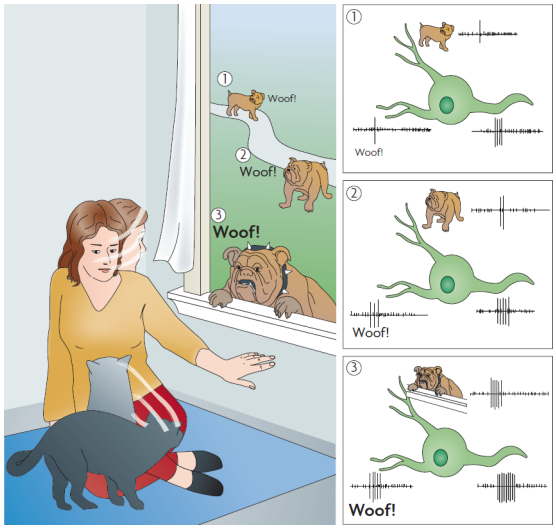


FIGURE 2.5 – Une illustration du renforcement multisensoriel. Dans trois situations, la femme et le chat entendent et voient un chien par la fenêtre. Dans la première, le chien est loin, les informations auditives et visuelles sont faibles. Le renforcement est par conséquent très fort ; la réponse neuronale bimodale dépasse la somme des deux réponses unimodales prises séparément (comportement *super-additif*). Au fur et à mesure que le chien se rapproche, le son et l'image deviennent facilement identifiables, jusqu'à la troisième situation parfaitement claire. Le renforcement s'affaiblit ; la réponse bimodale, bien que toujours plus forte que chacune des réponses unimodales séparées (i.e., il y a toujours renforcement), est néanmoins inférieure à leur somme (comportement *sous-additif*). Image empruntée à [STEIN et STANFORD, 2008].

ces variations tient au type de stimuli, à leur combinaison, à l'impact de leurs propriétés physiques sur l'organisme, et à leur temps d'apparition. [Meredith, Nemitz et Stein, 1987] insiste particulièrement sur la synchronisation temporelle des stimuli. L'attention du sujet serait aussi déterminante (voir [Wright et Ward, 2008] pour une revue complète de la question). Mais si ces considérations permettent de mieux comprendre nos mécanismes d'intégration sensorielle au plus bas niveau de la perception, il est encore difficile de quantifier précisément dans quelle mesure audition et vision contribuent à former le percept auditivo-visuel.

Plusieurs études ont tenté de formaliser cette contribution, en s'appuyant particulièrement sur la perception d'incohérences que l'on observe au niveau comportemental. L'incohérence peut être de plusieurs types : spatiale (e.g., un son et une image synchronisés et sémantiquement cohérents mais localisés à des endroits différents), temporelle (e.g., un son et une image localisés au même endroit et sémantiquement cohérents mais désynchronisés temporellement), ou sémantique (e.g., un son et une image coïncidents dans l'espace et le temps mais sans lien sémantique l'un avec l'autre), ou un mélange des trois. Lorsqu'elle engendre une interprétation déformée de la réalité, elle est souvent réifiée comme une illusion perceptive, ou « effet ». Nous en donnons trois exemples ici. L'effet ventriloque est le premier, résultat de la perception d'une discongruence spatiale entre un son et une image [Howard et Templeton, 1966 ; Chen et Vroomen, 2013]. Il décrit la tendance à percevoir deux stimuli proches dans l'espace, mais distincts, comme localisés au même endroit. Un autre exemple, lié cette fois à une incohérence temporelle entre son et image, est l'effet double-flash [Shams, Kamitani et Shimojo, 2000] selon lequel un sujet qui se voit présenter un flash lumineux accompagné de deux clics sonores rapporte avoir vu deux flashs lumineux. Enfin, un dernier exemple lié à l'incohérence sémantique entre son et image est l'effet McGurk [McGurk et MacDonald, 1976] : lorsqu'on présente à un sujet un interlocuteur dont la voix émet le son « ba » avec ses lèvres qui prononcent « ga », il perçoit l'objet audiovisuel « da ».

À l'inverse, l'interprétation d'un percept auditivo-visuel peut être renforcé dans une direction par une des deux modalités quand l'autre prête à confusion, comme par exemple avec l'effet rebond, introduit par [R. Sekuler, A. B. Sekuler et Lau, 1997]. Dans leur expérience, deux disques visuels suivent deux trajectoires rectilignes opposées, amenées à se rencontrer. Le moment du croisement peut s'interpréter de deux façons : soit les disques se traversent et continuent leur chemin, soit ils se cognent et repartent dans la direction opposée. Dans un cas purement visuel, la majorité des sujets voient des disques qui se traversent. Mais en ajoutant un clic sonore au moment de croisement, la tendance s'inverse et les sujets voient alors des disques qui rebondissent.

Pour en revenir à la formalisation, dans tous ces cas évoqués, comment se combinent nos perceptions auditive et visuelle ? Dans [Welch et Warren, 1980], les auteurs évoquent une hypothèse de pertinence de la modalité (*modality appropriateness hypothesis*) qui soutient que les incohérences multisensorielles sont résolues à la faveur de la modalité dominante selon la dimension étudiée : la vision domine l'audition pour des tâches spatiales, tandis que l'audition domine la vision pour des tâches temporelles. Mais si cette règle capture sans doute l'essentiel des mécanismes perceptifs, elle n'est pourtant pas toujours valable. Par exemple dans [Alais et Burr, 2004], les auteurs montrent que lorsque la qualité de l'image est dégradée, l'attraction spatiale

de l'effet ventriloque peut être renversée, exercée par le stimulus auditif sur le visuel. Pour quantifier et formaliser davantage ces subtilités, des modèles mathématiques plus récents présentent la perception d'un objet multimodal comme une somme pondérée des signaux unimodaux. Cette pondération se fait essentiellement via des modèles de probabilité bayésiens et des méthodes d'estimation de paramètres [Ernst et Banks, 2002 ; Ernst et Bülthoff, 2004 ; Kersten, Mamassian et Yuille, 2004 ; Burr et Alais, 2006]. Ils permettent alors de mieux décrire notre perception telle qu'elle semble fonctionner dès le niveau neuronal, en intégrant les variations dues au contexte, aux types de stimuli, à l'attention du sujet, etc.

2.2.2 Le cas de l'effet ventriloque, méthodes de mesure et résultats

Dans notre cas, nous ne cherchons pas à modéliser les mécanismes d'intégration multisensorielle mais à cerner les effets perceptifs qui peuvent intervenir à l'utilisation d'une application audiovisuelle sur mobile. En particulier, nous nous intéressons à ceux qui peuvent émerger par l'ajout du binaural, dont la particularité est d'enrichir la représentation spatiale de la scène sonore. L'effet ventriloque nous paraît être l'effet principal répondant à ce critère de sélection.

En effet, l'effet ventriloque nous permet de questionner l'importance de la coïncidence spatiale entre son et image. Du point de vue applicatif, il n'est pas nécessairement considéré comme indésirable. En témoignent [Moeck et al., 2007], qui l'utilisent pour alléger le calcul de positionnement de sources auditives dans un jeu vidéo en les groupant dans une scène virtuelle. Les imprécisions spatiales qui en résultent sont compensées perceptivement par l'effet ventriloque. Dans le même ordre d'idée, [Bruijn et Boone, 2003] et [André et al., 2014 ; Mannerheim, 2011] évoquent l'exploitation de cet effet, le premier dans un contexte de vidéoconférence avec un écran 2D, les seconds avec une image 3D stéréoscopique (typique de celles utilisées dans certaines salles de cinéma), et les deux avec un système de rendu audio spatialisé WFS. Dans une application mobile, la localisation du visuel se limite à l'écran, de taille réduite. Un signal rendu en mono ou en stéréo sur casque étant localisé à l'intérieur de la tête, l'association entre son et image introduit déjà une décorrélation spatiale (constituée par la distance allant de la tête au smartphone). Pour autant elle ne pose pas de problèmes en termes d'acceptation dans une utilisation quotidienne d'un téléphone mobile. Peut-être la raison tient au fait que la localisation à l'intérieur de la tête correspond à une non-spatialisation du point de vue de l'auditeur, comme le mentionne par exemple [Van der Burg, Olivers et al., 2008], et qu'une synchronisation temporelle efficace suffit à intégrer les deux modalités ([Van der Burg, Cass et al., 2010], synchronisation dont l'importance était déjà suggérée par [Meredith, Nemitz et Stein, 1987] au niveau neuronal). L'ajout du binaural constitue alors l'ajout d'une dimension spatiale pour le son, qui n'existait pas jusqu'alors au casque, et par conséquent une dissociation entre son et image compensée ou non par l'effet ventriloque.

Les premières observations scientifiques de l'effet ventriloque datent de la fin du XIX^{e} et du début du XX^{e} siècle. Des expériences consistant à porter un dispositif pour inverser la vision de l'œil droit avec celle de l'œil gauche [Stratton, 1897 ; Ewert, 1930], ou l'audition de l'oreille droite avec celle de l'oreille gauche (le pseudophone, voir Figure 2.6) [Young, 1928 ; Willey, Inglis et Pearce, 1937] ont été menées dans le but

FIGURE 2.6 – Le pseudophone de Young, un dispositif permettant d'inverser l'audition de l'oreille droite avec celle de l'oreille gauche. Image tirée de [VURPILLOT, 1963].

d'observer leurs effets sur la perception. Des observations ont été rapportées de sujets percevant certaines sources auditives comme localisées à l'endroit du visuel, malgré l'inversion. Cet effet, d'abord observé de façon qualitative, a depuis fait l'objet de nombreuses mesures quantitatives, dans le but de mettre en lumière son influence sur notre perception. Nous en présentons ici quelques unes, classées par ordre chronologique.

Une des premières études sur la question est présentée dans [WITKIN, WAPNER et LEVENTHAL, 1952], dans laquelle le sujet se trouve dans une cabine insonorisée, face à un interlocuteur qui lui parle à travers une fenêtre. Le son lui parvient aux oreilles par des écouteurs reliés à des tuyaux en laiton dont les expérimentateurs font progressivement varier la taille pour faire varier son ITD, simulant ainsi au sujet un mouvement de la source sonore sur le plan horizontal (mouvement en azimut). Les sujets doivent lever la main dès qu'ils perçoivent un déplacement du son au delà du centre. Les résultats révèlent la présence d'un effet ventriloque pour un décalage du son allant en moyenne jusqu'à 33 ° .

Dans une autre expérience [JACKSON, 1953], les sujets sont installés face à un mur en arc de cercle sur lequel ont été fixées sept bouilloires, réparties à intervalles fixes, desquelles s'échappe silencieusement de la fumée. Derrière le mur, sept autres bouilloires sifflantes sont cachées à la même position. À chaque session est présentée une bouilloire « auditive » qui siffle en même temps qu'une bouilloire « visuelle » qui fume. Le sujet doit identifier à chaque fois de quelle bouilloire il entend le son et de quelle bouilloire il voit la fumée. Ici, l'effet ventriloque en horizontal, testé pour des positions discrètes des stimuli, avoisine également les 30 ° .

Dans [CHOE et al., 1975], de façon analogue à [JACKSON, 1953], les auteurs proposent une expérience dans laquelle le sujet fait face à des sources visuelles (des diodes lumineuses) accrochées à un mur derrière lequel sont aussi fixées des sources auditives (des haut-parleurs). La différence avec [JACKSON, 1953] réside dans le fait que le sujet ne doit pas localiser une source, mais simplement dire si oui ou non il perçoit une

discordance spatiale entre la diode qui s'allume et le son qu'il entend. Le décalage du son par rapport au visuel est soit de 0 °, soit de 11 ° en azimut à gauche ou à droite. Les résultats indiquent une forte propension à considérer les stimuli alignés dans tous les cas. Mais plutôt que de l'interpréter comme un effet perceptif, les auteurs utilisent un modèle de décision statistique pour statuer que cela est davantage dû à un biais de décision, causé par la synchronisation temporelle du son et de l'image. Cette assertion est toutefois critiquée plus tard par [Bertelson et Radeau, 1976].

Dans [Thurlow et Jack, 1973], l'effet est mesuré sur une télévision couplée à un haut-parleur caché, les deux à une position fixe. Les sujets doivent indiquer avec un chronomètre le temps pendant lequel ils perçoivent un alignement spatial entre son et image. Les auteurs s'appuient donc davantage sur la durée de coïncidence spatiale perçue que sur une détection ponctuelle. Les résultats montrent une perception d'alignement la majorité du temps pour la plupart des sujets, pour un décalage spatial en élévation allant jusqu'à 195 ° entre le son et l'image (le haut parleur étant placé derrière le sujet et la télévision devant), alors que le son est pourtant correctement localisé en l'absence d'image. Avec un haut-parleur décalé horizontalement, l'alignement est ressenti par une majorité de sujets et la majorité du temps pour un décalage de 20 °, mais il n'est presque pas ressenti à 60 °. Ces résultats indiquent une nette différence de notre capacité à localiser entre le plan horizontal et le plan vertical. Dans [Jack et Thurlow, 1973], les mêmes auteurs montrent dans des conditions expérimentales similaires que l'effet ventriloque agit toujours pour un décalage angulaire de 30 ° en azimut entre son et image, mais décroît significativement à 40 ° et au delà. Par ailleurs, une source sonore placée exactement derrière le sujet (angle d'azimut égal à 180 °) est également fortement perçue comme corrélée avec le visuel placé devant, alors que c'est moins le cas pour des sources auditives placées à 160 ° ou 140 °. Ces résultats en élévation suggèrent que l'effet ventriloque peut être facilement provoqué dans le plan médian, probablement à cause de l'absence d'ITD et d'ILD.

De la même façon dans [Radeau et Bertelson, 1977], le sujet doit indiquer le temps pendant lequel il perçoit une fusion entre son et image, le son étant retransmis sur un haut-parleur parmi trois situés à 0 °, 20 ° à gauche et 20 ° à droite du sujet, légèrement en contrebas (sous une table), et l'image étant retransmise sur un écran aligné avec un des deux haut-parleurs à 20 °. Plusieurs types de stimuli sont présentés, et les résultats indiquent une fusion perçue en moyenne jusqu'à 77% du temps d'exposition pour les stimuli sémantiquement cohérents et temporellement synchronisés.

Dans [Komiyama, 1989], les sujets délivrent leur sentiment d'incohérence spatiale ressenti en étant exposés à un visuel retransmis sur une télévision face à eux, et à un stimulus auditif retransmis sur une des dix enceintes qui flanquent la télévision à sa droite (décalage en azimut). Sept sont réparties entre 0 ° et 45 °, une est positionnée à 90 °, une à 135 ° et une à 180 °. Dans une première version de l'expérience, le stimulus est constitué d'une femme qui lit un journal à voix haute, puis dans un second temps c'est un programme de télévision réel, une chanson pop interprétée en concert par une femme. Pour noter son sentiment d'incohérence, le sujet doit utiliser l'échelle de notation tirée des recommandations de l'Union Internationales de Télécommunications (UIT) [ITU-R, 2002] : A. Imperceptible - B. Perceptible mais non dérangeant - C. Légèrement dérangeant - D. Dérangeant - E. Très dérangeant. Les résultats indiquent pour les deux types de contenu que plus le décalage spatial entre son et image est élevé, moins les scores sont bons, à l'exception de l'enceinte à 180 ° (arrière du sujet

dans le plan médian). En moyenne, un décalage allant jusqu'à 20 ° est accepté par des sujets non-experts.

Dans [Lewald, Ehrenstein et Guski, 2001], les auteurs introduisent le mouvement. L'effet ventriloque est mesuré dans le cas d'un visuel qui se déplace en azimut. Ce visuel est la lumière d'un laser projeté sur un écran incurvé en demi-cercle face au sujet. Le son est quant à lui un train de sinus retransmis sur une enceinte parmi trois, placées à 10 °, 12 ° et 14 ° à droite du plan médian, et cachées au sujet. Le laser se déplace d'un mouvement continu, partant d'une position en périphérie de l'écran, croise les haut-parleurs, puis les dépasse. Selon la méthode présentée, appelée point d'alignement spatial subjectif (ou PSSA pour *Point of Subjective Spatial Alignment*), le sujet doit presser un bouton au moment où il perçoit un alignement entre son et image. Les résultats indiquent un point d'alignement d'environ 3 ° avant l'alignement réel. Par la suite [Lewald et Guski, 2003], les auteurs trouveront des résultats similaires avec un protocole très différent impliquant des stimuli fixes : un sinus rendu sur haut-parleur (parmi 11, placés sur un demi-cercle faisant face au sujet) et un flash lumineux émis par une diode, positionnée en dessous du haut-parleur central. Les sujets doivent évaluer la « vraisemblance selon laquelle son et image sont alignés spatialement », en utilisant une échelle de notation à 9 valeurs. Les résultats montrent une chute significative des notes au-delà d'un décalage de 4 ° à gauche ou à droite entre le son et l'image. Pour autant, la sensation d'alignement spatial obtient des notes correctes jusqu'à 11 °. Fait intéressant, avec le même protocole, mais en demandant au sujet d'évaluer « la vraisemblance selon laquelle son et image partagent une cause commune » (i.e., s'il y a un phénomène de causalité qui relie les stimuli visuel et auditif), la tolérance s'élargit à 14 °.

Dans [Bruijn et Boone, 2003], les auteurs évaluent l'écart spatial en azimut acceptable pour un visuel rendu sur un écran couplé à un rendu sonore en WFS, dans un contexte de vidéoconférence. Le sujet observe trois hommes debout, statiques, représentés à l'échelle 1 :1 sur l'écran et positionnés à des profondeurs différentes de l'image. Les stimuli auditifs sont des voix d'homme qui lisent un texte en continu, positionnées chacune à l'endroit d'un des visuels, en respectant leur profondeur réelle (et malgré la mise à plat sur l'écran de l'image), ce qui crée un décalage perceptif entre son et image, essentiellement en azimut. Le sujet écoute une des trois voix choisie au hasard, doit identifier de quel visuel elle provient et noter son degré de gêne ressenti en utilisant l'échelle de notation UIT. Les auteurs estiment que dans ces conditions expérimentales, l'effet ventriloque est efficace jusqu'à 11 ° de décalage entre son et image. Dans des conditions expérimentales quasiment identiques, [Melchior, Brix et al., 2003] trouvent un décalage acceptable entre 5 ° et 7 °.

Dans la continuité de ces deux dernières expériences, [Melchior, Fischer et Vries, 2006] évaluent l'écart spatial accepté en azimut entre un son rendu en WFS et une image 3D projetée via des lunettes de réalité augmentée. Toujours avec le même protocole d'évaluation (noter le niveau de gêne du décalage perçu avec une échelle de notation UIT), les auteurs évaluent l'effet ventriloque efficace jusqu'à 8 °.

Dans [Nguyen et al., 2010], les auteurs reprennent le protocole du PSSA introduit par [Lewald, Ehrenstein et Guski, 2001] dans un contexte de réalité virtuelle (scène 3D visionnée sur un écran avec des lunettes 3D) couplée à du son binaural. C'est à notre connaissance la première étude à mesurer l'effet ventriloque avec du son binaural. Ici

les HRTF sont systématiquement individualisées. Dans une première occurrence de l'expérience, les stimuli sont composés d'un visuel d'hélicoptère qui bouge (allant du centre vers la périphérie du champ de vision et inversement), et d'un son de rotor fixe décalé de 12 ° en azimut par rapport à la position centrale. Les sujets doivent presser un bouton dès qu'ils perçoivent un alignement entre le son et l'image. Les résultats indiquent une perception d'alignement décalée par rapport à l'alignement objectif. Mais pour les deux types de trajectoires, l'alignement est perçu au delà des 12 °. Pour les trajectoires de la périphérie vers le centre, ce résultat est intuitif : le sujet perçoit un alignement légèrement avant que l'image ne croise effectivement le son. Mais pour les trajectoires du centre vers la périphérie, cela signifie que l'alignement est perçu en retard, après seulement que les deux sources se soient croisées. Aucune interprétation n'est fournie concernant cette différence entre les trajectoires. En revanche, l'étroitesse de la fenêtre d'intégration (environ 2.5 °, entre 14.1 ° et 16.6 °) obtenue est expliquée par la complexité des stimuli comparée aux autres expériences, ainsi que leur caractère continu. Dans une deuxième version de l'expérience, c'est le son qui bouge et le visuel qui est statique. Les résultats donnent une fenêtre d'intégration de taille similaire (2.5 °, entre 9.1 ° et 11.6 °), mais cette fois-ci décalée vers le centre par rapport au visuel (ce qui est considéré comme un résultat similaire à l'expérience d'avant, les deux modalités étant en situation inversée).

Dans [Werner, Liebetrau et Sporer, 2013], les auteurs se focalisent sur le décalage spatial en élévation entre un son rendu en binaural individualisé et une diode lumineuse. Le stimulus auditif est un son de saxophone ou un bruit blanc, placé à 0 ° ou 25 ° d'élévation en face du sujet (0 ° d'azimut). La diode lumineuse est sélectionnée parmi 10 diodes espacées régulièrement de -10 ° à 35 °, sur un demi-cercle vertical face au sujet. La même configuration décalée à 30 ° en azimut est également testée. Toutes les combinaisons son-image sont présentées au sujet, qui doit dire s'il perçoit le son au dessus, en dessous ou dans le plan du visuel. Les résultats indiquent une tolérance du sujet à des décalages entre 8 ° et 9 ° lorsque la source sonore est à élévation 0 °, et jusqu'à 17 ° lorsque la source sonore est élevée à 25 °. Ces résultats sur l'effet ventriloque en élévation, nettement plus faibles que dans [Thurlow et Jack, 1973], peuvent en partie s'expliquer par le fait que les sujets sont présentés comme très accoutumés aux tests d'écoute.

Dans [André et al., 2014], l'effet ventriloque est mesuré sur un système de rendu sonore WFS allié à une image projetée en 3D sur un écran (visionnée avec des lunettes passives), pour plusieurs positions du sujet (variation de la localisation de la source visuelle en azimut) et un son projeté à différentes positions (variation de la localisation de la source auditive en azimut). Le stimulus est constitué d'un homme dans un appartement, rendu à l'échelle 1 :1 via l'écran 3D, qui prononce des phrases. Les sujets passent l'expérience par trois, chacun alternativement placé à un endroit différent. Pour chaque position du son, les sujets doivent indiquer en pressant un bouton si oui ou non ils perçoivent la voix comme cohérente avec le positionnement de l'homme. Les résultats révèlent une intégration auditivo-visuelle significative jusqu'aux alentours de 19 ° de décalage perçu entre son et image.

Enfin dans [Hendrickx et al., 2015], les auteurs mesurent l'effet ventriloque pour des décalages simultanés en azimut et en élévation. Le stimulus visuel est un visage qui prononce une phrase, rendu en 3D sur l'écran d'un vidéoprojecteur face au sujet qui porte des lunettes 3D actives. Le son est la phrase qui correspond au visuel, rendue

sur une des 28 enceintes réparties sur un quart de sphère (supérieur droit) centré sur le sujet. À chaque essai, le sujet doit indiquer si oui ou non la voix et la bouche de l'acteur lui paraissent provenir de la même direction. Les résultats indiquent une forte disparité entre les sujets : pour une variation en azimut uniquement, un effet ventriloque qui opère pour des décalages allant de 7 ° à 21 ° ; pour des variations en élévation uniquement (sur le plan médian), un effet ventriloque plus large en moyenne, mais également plus disparate, qui varie pour un décalage allant de 19 ° à 137 °. Pour des décalages conjoints en azimut et en élévation, les valeurs sont intermédiaires, et le pouvoir d'attraction de l'image sur le son augmente d'autant plus qu'on s'approche du plan médian. Par ailleurs, les variations d'élévation semblent avoir moins d'impact sur l'effet ventriloque que les variations d'azimut, et ce même pour des trajectoires essentiellement verticales.

Pour résumer ce tour d'horizon chronologique, nous pouvons dire que l'effet ventriloque a été mis en évidence dans des situations variées. Nous ajoutons sans rentrer dans les détails que d'autres études avec des méthodes d'évaluation similaires font également état d'un effet ventriloque en profondeur, sous l'appellation de *proximity image effect* [Gardner, 1968 ; Mershon et al., 1980 ; Pavel Zahorik, 2003 ; Agganis, Muday et Schirillo, 2010 ; Bowen et al., 2011 ; Hládek et al., 2013]. Dans le cadre de cette thèse, nous mettons volontairement de côté cette question de profondeur et en particulier celle de la perception acoustique en champ proche (distance de la source à l'auditeur inférieure à 1 m). Elle nécessiterait de faire intervenir un cadre théorique à part entière et un déploiement technique (moteur de synthèse binaurale en champ proche, jeux de HRTF enregistrés en champ proche) dont nous ne disposons pas. Il conviendra cependant de l'aborder dans une étude ultérieure.

Pour le reste, la tendance principale qui émerge est que l'effet ventriloque agit quand les stimuli auditifs et visuels entretiennent une certaine relation spatiale (en étant proches l'un de l'autre principalement, ou en étant tous deux sur le plan médian), et décroît au fur et à mesure qu'on brise cette proximité. Les limites de l'attraction d'une modalité sur l'autre varient beaucoup d'une publication à l'autre (des décalages acceptés entre son et image allant de 2 ° à plus de 30 ° en azimut, selon les publications !). De nombreux facteurs d'influence ont été montrés comme étant déterminants pour mesurer l'effet ventriloque et entrent probablement en ligne de compte : la synchronisation temporelle entre les deux stimuli [Choe et al., 1975 ; Radeau et Bertelson, 1977 ; Lewald, Ehrenstein et Guski, 2001 ; Slutsky et Recanzone, 2001 ; Wallace et al., 2004], leur cohérence sémantique [Jackson, 1953 ; Thurlow et Jack, 1973 ; Radeau et Bertelson, 1977 ; Warren, Welch et McCarthy, 1981], leur qualité [Alais et Burr, 2004], l'attention du sujet [Spence et Driver, 2000 ; Talsma et al., 2010 ; Hendrickx et al., 2015] (mais pas l'attention visuelle selon [Bertelson, Vroomen et al., 2000 ; Vroomen, Bertelson et De Gelder, 2001 ; Lewald, Ehrenstein et Guski, 2001]), l'expérience d'écoute du sujet [Komiyama, 1989], l'acoustique de la salle [Jackson, 1953 ; Thurlow et Jack, 1973], et sans doute d'autres encore (par exemple l'influence du niveau de bruit ambiant est discutée par [André et al., 2014], bien que non-significative dans leur étude).

Toutefois, dans la mesure du possible, lorsque les expériences étaient déclinées selon divers paramètres, nous avons essayé de toujours présenter les résultats obtenus dans des conditions expérimentales les plus similaires possibles les unes aux autres : stimuli synchronisés, sémantiquement cohérents, de bonne qualité, attention du sujet

maximum, et expérience d'écoute non experte.

Peut-être qu'un autre facteur d'influence explique en partie ces différences de résultats, c'est la nature du protocole. Au-delà des disparités de matériel dues aux grands écarts de temps qui séparent parfois les expériences, nous dégageons quatre grands types de protocoles expérimentaux, que nous résumons de la façon suivante :

— le sujet doit identifier parmi plusieurs choix la source visuelle correspondant au son entendu ([Jackson, 1953]) ;

— le sujet doit rapporter le temps pendant lequel il perçoit une fusion entre un son et une image décalés spatialement ([Thurlow et Jack, 1973 ; Jack et Thurlow, 1973 ; Radeau et Bertelson, 1977]) ;

— le sujet doit dire si oui ou non il perçoit une fusion pour des décalages variés entre son et image [Choe et al., 1975 ; Hendrickx et al., 2015]. Dans une version plus sophistiquée, le oui/non est remplacé par une échelle de notation issue de l'UIT [Komiyama, 1989 ; Bruijn et Boone, 2003 ; Melchior, Brix et al., 2003 ; André et al., 2014] ou personnalisée [Lewald et Guski, 2003 ; Werner, Liebetrau et Sporer, 2013] ;

— le sujet doit indiquer le moment où il perçoit le point d'alignement ou le point de rupture spatial entre un son et une image lorsqu'un des deux se déplace et l'autre reste fixe [Witkin, Wapner et Leventhal, 1952 ; Lewald, Ehrenstein et Guski, 2001 ; Nguyen et al., 2010].

Remarquons que dans tous les cas, les protocoles expérimentaux reposent sur la perception du sujet et le sentiment d'unicité spatiale qu'il ressent alors même que son et image sont décalés dans l'espace. D'autres expériences existent, impliquant des tâches de localisation pour lesquelles le sujet doit indiquer précisément où il perçoit un son, alors que l'image correspondante est décalée. Ces expériences montrent en général qu'il y a un effet d'attraction spatiale de l'image sur le son [Pick, Warren et Hay, 1969 ; Bertelson et Radeau, 1981 ; Bertelson et Aschersleben, 1998 ; Alais et Burr, 2004 ; Wallace et al., 2004]. Pour autant, si cette attraction fait état d'un décalage perceptif du son, elle n'implique pas nécessairement un sentiment de fusion entre les deux stimuli. En ce sens, [Bertelson et Radeau, 1981] énoncent que cet effet ne correspond pas tout à fait à l'effet ventriloque, et s'y réfèrent comme d'un biais inter-sensoriel. Lui-même et d'autres [Wallace et al., 2004] analysent d'ailleurs la relation qu'entretiennent les deux, relation proche malgré des différences. À noter enfin que plusieurs travaux utilisent le terme de ventriloquisme indifféremment pour l'un ou l'autre effet (à commencer par Berteslon lui-même dans [Bertelson et Aschersleben, 1998], mais aussi dans [Alais et Burr, 2004 ; Wallace et al., 2004] par exemple).

2.2.3 Conclusion

Dans ce chapitre, l'état de l'art nous montre que la perception spatiale d'un son binaural n'est pas univoque : d'une part elle est soumise à l'influence des paramètres du système de rendu, en particulier le caractère individualisé ou non des HRTF, et d'autre part elle est liée à son association avec la scène visuelle et au potentiel effet ventriloque qui en résulte. Par conséquent, à la question apparemment simple « où puis-je placer une source auditive binaurale pour qu'elle soit perceptivement associée

à son correspondant visuel ? », la réponse n'est en fait pas immédiate. De plus, malgré de nombreuses pistes de réflexion, cet état de l'art ne nous fournit pas de résultats explicites 1) sur la possibilité de se passer de HRTF individuelles dans un cadre audio-visuel, ni 2) sur l'intégration auditivo-visuelle spatiale, dont les expériences mènent à des résultats très disparates, tandis qu'aucune ne porte spécifiquement sur un son binaural couplé à un visuel sur terminal mobile. Pourtant, nous réaffirmons l'importance de ces deux points dans le cadre de notre thèse.

Dans le chapitre suivant, nous présentons une expérience dont le but est de nous fournir des éléments de réponse sur ces questions perceptives. Le choix de la méthode est discuté, puis nous présentons les détails de l'expérience, son déroulement et les résultats obtenus.

Chapitre 3

Mesure de la fenêtre d'intégration auditivo-visuelle entre son binaural et visuel sur mobile

3.1 Introduction

Dans le chapitre précédent, nous avons recensé les travaux menés sur l'individualisation du binaural et sur l'intégration spatiale auditivo-visuelle. Nous avons également exposé en conclusion certains manquements de cet état de l'art. Nous présentons ici une expérience dont le but est de nous informer sur notre perception spatiale d'un son binaural couplé à un visuel sur terminal mobile.

3.1.1 Enjeux de l'expérience

Dans un scénario typique d'utilisation grand public du binaural avec un smartphone, nous avons vu que l'individualisation des HRTF n'est pas envisageable en l'état. Mais alors, dans le cas non-individualisé, y a-t-il plus de difficultés à intégrer sensoriellement un visuel avec une source sonore décorrélée spatialement ? Aucune publication présentée dans le chapitre précédent ne répond à la question. Certains travaux suggèrent que l'ajout du visuel pourrait gommer la différence de perception entre des HRTF individualisées et des HRTF standard. Par ailleurs, dans la continuité de ce raisonnement, nous avons présenté l'effet ventriloque qui décrit l'influence réciproque qu'entretiennent des sources sonores et visuelles sur la perception de leurs positions respectives. Il pourrait donc être un de ces facteurs propres à relativiser l'impact de l'un ou l'autre type d'HRTF.

Mesurer l'effet ventriloque nous permet de connaître le découplage spatial possible entre une image de taille réduite et un son binaural « en dehors » de l'écran, sans pour autant troubler leur fusion perçue. Nous envisageons deux avantages possibles à utiliser ce découplage : premièrement, agrandir spatialement la scène sonore, au-delà de la scène visuelle, dans le but d'améliorer l'expérience utilisateur ; deuxièmement, s'affranchir du problème de la disparité spatiale entre la scène visuelle dont on ne

connait pas l'emplacement (on ne sait pas comment l'utilisateur tient le téléphone dans ses mains) et une scène sonore dont les sources sont placées à l'avance.

La mesure des capacités d'intégration auditivo-visuelle sur mobile dans le cas du binaural et l'évaluation de l'influence de l'individualisation des HRTF s'imposent donc comme préalable à toute considération sur l'apport du binaural à l'expérience utilisateur, ayant pour objectif de nous informer sur la perception d'un individu. Notons bien que nous ne cherchons pas à répondre définitivement aux questions de l'individualisation et de l'effet ventriloque. Cette expérience se limite bien évidemment au contexte du son binaural sur mobile.

3.1.2 Choix de la méthode de mesure : le PSSA

Nous souhaitons mettre au point une expérience pour mesurer les limites de l'intégration auditivo-visuelle lorsqu'une source binaurale est décorrélée spatialement de son correspondant visuel sur mobile. Nous voulons également savoir si cette intégration a pour facteur d'influence le type des HRTF, individualisées ou non. Dans cette sous-section nous allons discuter du protocole qui nous semble le plus adapté pour procéder à cette mesure. En guise d'avant-propos, rappelons qu'un des points importants de cette thèse (discuté dans l'introduction) est de mettre les sujets dans une situation la plus proche possible d'une utilisation réelle et quotidienne d'un smartphone. Cette volonté peut néanmoins être contrebalancée par la nécessité d'avoir un protocole reproductible d'un sujet à l'autre, pour obtenir des résultats agrégeables et statistiquement exploitables. Cette double préoccupation dictera certains choix expérimentaux, dont nous discuterons au cas par cas lorsqu'ils se présenteront dans ce chapitre.

Comme point de départ, l'état de l'art offre plusieurs pistes de réflexion. Nous nous inspirons de la section sur l'effet ventriloque du chapitre précédent, dont les mesures perceptives impliquent systématiquement des sources auditives et visuelles décorrélées spatialement. Nous avons déjà distingué quatre catégories de protocoles : 1) le sujet identifie celle des sources visuelles qui correspond spatialement au son entendu ; 2) le sujet reporte le temps pendant lequel il perçoit une fusion entre un son et une image statiques, décalés spatialement ; 3) le sujet indique si oui ou non il perçoit une fusion pour des décalages variés entre son et image ; et enfin 4) le sujet indique le moment où il perçoit le point d'alignement ou le point de rupture spatial entre un son et une image avec un des deux stimuli en mouvement et l'autre statique.

La première catégorie, essentiellement utilisée par [Jackson, 1953], est discutée par [Radeau et Bertelson, 1977]. Pour rappel, il s'agissait de montrer au sujet un ensemble de bouilloires dont une seule fumait et une autre (cachée) sifflait. Le sujet devait identifier de quelle bouilloire provenait le sifflement. [Radeau et Bertelson, 1977] argumentent qu'un tel procédé incite le sujet à choisir la bouilloire qui fume, quand bien même il n'y aurait pas de sentiment de fusion. Savoir simplement que, dans la vie de tous les jours, la bouilloire qui fume est aussi censée siffler suffit à orienter son choix, et ce même en dépit d'une information sensorielle contradictoire. Ce biais de décision est appuyé par une expérience de [Choe et al., 1975], qui ajoute que la simple synchronisation temporelle des deux stimuli, au-delà de leur cohérence sémantique, suffit à biaiser le choix du sujet, y compris pour la troisième catégorie de

protocoles. Toutefois, [Bertelson et Radeau, 1976] remettront méthodiquement en cause cette dernière affirmation. Par ailleurs, [Radeau et Bertelson, 1977] montrent plus d'indulgence pour le deuxième type de protocole, intronisé par [Thurlow et Jack, 1973], qui mesure le temps de fusion perçue. La quatrième catégorie quant à elle n'a pas fait l'objet de discussion à notre connaissance.

Cependant, un élément déterminant nous fait pencher pour cette dernière, c'est le fait qu'un des deux stimuli soit en mouvement. Dans les applications mobiles susceptibles d'être rendues en binaural (films, jeux vidéo, applications de vidéoconférence, etc.), les sources auditives et visuelles sont souvent amenées à bouger. Dans un souci de vraisemblance, il serait intéressant d'introduire ce mouvement dans notre expérience. En outre, [Wightman et Kistler, 1999] ont montré que la localisation auditive est favorisée par les indices dynamiques, en particulier pour la synthèse binaurale. Ce caractère dynamique peut être obtenu de deux façons, par le mouvement de la tête du sujet ou en faisant bouger les sources. C'est ce dernier cas qui nous intéresse. Les mouvements permettent notamment de résoudre les confusions avant-arrière, courantes avec le binaural. Dans leurs expériences, [Wightman et Kistler, 1999] testent la localisation de sources réelles (des haut-parleurs placés autour du sujet) et virtuelles (rendues sur casque en binaural avec des HRTF individualisées). Ils montrent dans les deux cas que lorsque la source sonore est en mouvement, les confusions ne peuvent se résoudre que si le sujet est informé de la direction de la trajectoire. Il apparaît donc important d'être explicite sur cette information dans les instructions fournies au sujet.

Par ailleurs, dans les différentes expériences sur l'effet ventriloque avec mouvement d'un stimulus, seuls [Witkin, Wapner et Leventhal, 1952] et [Nguyen et al., 2010] font bouger le stimulus auditif. Nguyen présente aussi une version de l'expérience où le visuel est en mouvement avec une source sonore fixe, pour des résultats comparables. Néanmoins, dans son manuscrit de thèse [Nguyen, 2012, p. 219], Nguyen revient sur les deux résultats pour préciser que 90% des sujets ont trouvé la tâche d'alignement plus simple lorsque le son est en mouvement, plutôt que le visuel : suivre le mouvement du visuel tout en se concentrant sur la position statique de la source sonore semble plus difficile que de se concentrer sur le mouvement sonore avec le regard fixe. Dans notre cas, le mouvement du visuel plutôt que de la source sonore est de toute façon difficilement envisageable, à cause de la petite taille de l'écran.

Enfin, une des plus grosses différences entre l'expérience présentée par [Witkin, Wapner et Leventhal, 1952] et celle de [Nguyen et al., 2010] réside dans l'état initial des stimuli et la tâche demandée au sujet. Dans la première, le stimulus sonore est en premier lieu aligné avec le visuel, s'en éloigne progressivement, et le sujet doit indiquer à quel moment il perçoit la rupture entre les deux. Dans la seconde, c'est l'inverse, la trajectoire commence alors que le stimulus sonore est explicitement décorrélé du visuel, et le sujet doit indiquer quand il perçoit un alignement (méthode du point d'alignement spatial subjectif, ou PSSA). Cette différence, en plus de celles importantes de matériel, contribue peut-être à expliquer le grand écart entre les deux résultats : une fenêtre d'intégration de 66 ° pour la première, de 2.5 ° pour la seconde. Il est possible que démarrer par des stimuli déjà séparés rende plus difficile la perception d'une fusion. À l'inverse, partir dès le départ avec la perception d'une fusion favorise peut-être une détection de rupture plus tardive. Ainsi, et eu égard au caractère plus récent de l'expérience de Nguyen, à la proximité de certaines caractéristiques, comme l'emploi du binaural, et de ce fait à la possibilité de comparer qualitativement les

résultats, nous choisissons d'adopter la méthode PSSA.

Notre expérience consiste donc à proposer à des sujets un stimulus visuel fixe sur un écran de smartphone, et un stimulus sonore en mouvement horizontal, dont la position initiale est décorrélée du visuel. Le sujet doit indiquer le point où il perçoit un alignement ou une fusion entre les deux. Deux mouvements sonores opposés permettent de collecter deux PSSA et de définir ainsi les limites de la fenêtre d'intégration auditivo-visuelle, à l'intérieur de laquelle les sujets considèrent que les deux stimuli sont perçus comme confondus (autrement dit à l'intérieur de laquelle l'effet ventriloque est opérant). Plusieurs paramètres viennent faire varier ce protocole initial pour répondre aux questions amenées par le chapitre précédent (notamment celles de l'influence de l'individualisation des HRTF et de la différence de position en élévation entre l'écran et le stimulus auditif). Cette section n'a toutefois pas pour but de pousser plus en avant la présentation du protocole. Nous justifions simplement ici du choix général de la méthode à la lumière de l'état de l'art. Les détails de l'expérience, matériel, procédure, etc. seront exposés dans les sections suivantes.

3.1.3 Facteurs d'influence de l'effet ventriloque

Ajoutons un dernier point sur les facteurs d'influence de l'effet ventriloque. Nous avons vu dans le chapitre précédent qu'ils étaient nombreux, et tous les inclure comme paramètres rendrait l'expérience longue et fastidieuse. Encore une fois nous nous appuyons sur une utilisation vraisemblable du smartphone dans notre contexte d'application pour nous affranchir de la plupart d'entre eux. On suppose donc une synchronisation temporelle entre le son et l'image, une bonne qualité des stimuli (pas de dégradation volontaire) et une cohérence sémantique entre le visuel et l'auditif. En ce qui concerne le sujet, il aura une expérience d'écoute variable (non contrôlée), et une attention portée sur l'expérience (pas de détournement d'attention volontaire). Ce dernier point pourrait prêter à discussion, étant donné l'aspect perturbant de certains contextes d'utilisation d'un smartphone. Néanmoins, l'attention est elle-même sujette à des paramètres multiples (elle peut être sollicitée par l'une ou l'autre modalité sensorielle, délibérée ou automatique, via des stimuli simples ou complexes), pour des résultats variables sur l'intégration multisensorielle et l'effet ventriloque [Bertelson, Vroomen et al., 2000 ; Spence et Driver, 2000 ; Vroomen, Bertelson et De Gelder, 2001 ; Lewald, Ehrenstein et Guski, 2001 ; Talsma et al., 2010 ; Hendrickx et al., 2015]. Nous intéresser à l'effet de l'attention sur la fenêtre d'intégration constituerait une tâche à part entière. Nous prenons donc le parti de ne pas l'intégrer comme critère de variation, et de considérer uniquement une attention « optimale » du sujet.

3.2 Description de l'expérience et hypothèses

3.2.1 Matériel et logiciel

Nous utilisons un téléphone Motorola Nexus 6 XT1100 (Figure 3.1), dont l'écran mesure 5,96 pouces de diagonale, soit une taille de 132x74mm (ratio 16/9), pour une

FIGURE 3.1 – Le matériel utilisé : en haut à gauche le smartphone Motorola Nexus 6 ; en haut à droite le contrôleur bluetooth universel ; en bas le casque Sennheiser HD650.

résolution de 1440x2560px. L'écran est de grande taille par rapport à la moyenne[1], pour nous assurer de la bonne visibilité du stimulus visuel. Le système d'exploitation utilisé est Android 7.1.1.

Le son est retransmis via un casque Sennheiser HD650 (Figure 3.1), casque circum-aural (c'est-à-dire qui englobe complètement l'oreille). Nous avons opté pour un casque ouvert car il permet une circulation libre de l'air entre l'intérieur et l'extérieur des coques qui entourent les haut-parleurs, réduisant la sensation de fatigue sur une longue durée d'écoute (en comparaison d'un casque fermé, dont l'emprisonnement des membranes à l'intérieur des coques peut engendrer des surpressions de l'air). Un désavantage du casque ouvert est son absence d'isolation phonique, mais dans notre cas l'expérience se déroule dans une salle acoustiquement isolée. Par ailleurs, le choix du modèle HD650 s'appuie sur une décision consensuelle d'experts l'ayant utilisé lors du projet BiLi[2].

Le sujet indique l'alignement des deux stimuli à l'aide d'un contrôleur bluetooth universel (Figure 3.1), dont la latence a été testée pour ne pas excéder celle d'un clavier d'ordinateur classique.

Les stimuli auditifs et visuels sont présentés au sujet via une application développée avec le logiciel Unity, version 5.4.2f2 Personal, logiciel adapté au développement de jeu vidéo, idéal pour construire une scène audiovisuelle 3D avec un déplacement des objets.

1. La taille moyenne d'un écran de smartphone étant de 4,9, calcul basé sur 113 modèles, à l'aide du site http://screensiz.es/phone

2. un projet de recherche collaboratif, impliquant entre autres partenaires Orange, et dont les objectifs sont liés à l'acquisition, la restitution, la diffusion et l'évaluation du binaural http://www.bili-project.org/

FIGURE 3.2 – Les stimuli visuels : un halo lumineux à gauche ; un modèle 3D d'hélicoptère à droite.

3.2.2 Stimuli et organisation de la scène virtuelle

Présentation des stimuli audiovisuels

Dans le but de vérifier que la fenêtre d'intégration n'est pas trop sujette à l'influence du contenu sémantique du stimulus, nous choisissons de la mesurer pour deux types de stimuli audiovisuels : des stimuli sémantiquement neutres et des stimuli porteurs de sens. Dans ce dernier cas, en plus d'être porteurs de sens, notons bien qu'ils sont également sémantiquement cohérents, tels que l'avons déjà discuté. Par ailleurs, nous nous inspirons ici des travaux de Nguyen [NGUYEN, 2012] sur le choix des stimuli, toujours dans le but de comparer les résultats.

Les stimuli sémantiquement neutres sont composés d'un halo lumineux, généré directement sur le logiciel Unity (voir Figure 3.2), et d'une salve de bruits blancs, c'est-à-dire une série de bruits blancs de 50ms séparés par des silences de 10ms, généré via Matlab.

Les stimuli porteurs de sens et sémantiquement cohérents sont composés pour le visuel d'un modèle 3D d'hélicoptère (modèle d'un HH-65C Dauphin [3], voir Figure 3.2), dont les pales sont animées et tournent à vitesse constante. Le stimulus audio est un son de rotor d'hélicoptère, synthétisé avec le logiciel Ableton Live, échantillonné à 44100Hz [4]. Le son d'hélicoptère possède l'avantage d'être relativement riche sur le plan fréquentiel et de faciliter ainsi sa localisation [BLAUERT, 1997].

La scène virtuelle

Les positions des stimuli sonores et visuels sont ajustés dans Unity, dans une représentation 3D de la scène. En plus de ces objets, une caméra virtuelle est figurée. Elle est chargée de capturer l'image et de la retransmettre sur l'écran du téléphone.

Nous insistons ici sur la distinction entre la représentation de la scène visuelle dans Unity, la position relative des objets, les distances et les angles qui les séparent, et sa représentation « aplatie » à l'écran. Unity modélise le fonctionnement de la caméra de

3. téléchargé gratuitement sur la boutique en ligne de Unity, l'Asset Store : https://assetstore.unity.com/packages/3d/vehicles/air/hh-65c-dauphin-8128

4. téléchargé gratuitement sur la banque de sons Freesound : http://freesound.org/people/fridobeck/sounds/194250/

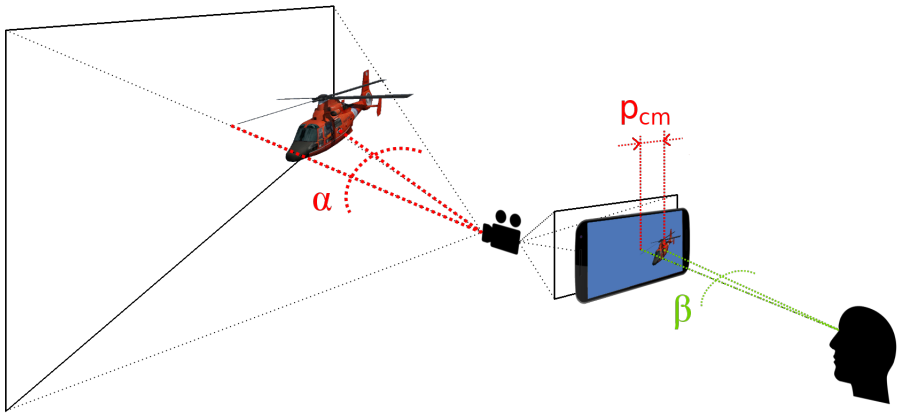


FIGURE 3.3 – Schéma du processus de représentation de la scène visuelle, de l'espace virtuel (à gauche, telle que représentée dans Unity) au monde réel (à droite, observée par le sujet).

façon réaliste, en restituant la déformation induite par les propriétés de sa lentille, qui permet de conserver un sens de la perspective et de la profondeur sur l'image aplatie (à noter que la caméra Unity aurait pu aussi capturer l'image de façon orthographique, c'est-à-dire sans conserver aucun sens de la perspective). Il est donc important de noter que les distances et angles entre les objets visuels dans la scène virtuelle ne sont pas conservés à l'identique à l'écran : ils subissent non seulement la déformation de la caméra, mais sont aussi transformés selon la taille et la résolution de l'écran du téléphone. Enfin, la perception de cette retransmission par le sujet dépendra également de sa distance au téléphone (voir Figure 3.3).

À partir de la position d'un objet visuel dans la scène virtuelle, nous calculons son angle en azimut observé par l'utilisateur par rapport au centre de l'écran. La première étape consiste à convertir la position de la source dans la scène virtuelle (fournie en coordonnées cartésiennes $[x; y; z]$), en position en pixels sur l'écran. Sans rentrer dans les détails, Unity utilise l'équation suivante, qui tient compte des caractéristiques de la caméra :

$$p_{h_{px}} = 0.5 * R_h * (\frac{0.97428x}{z}) + 1,$$

avec R_h la résolution horizontale de l'écran en pixels. Le résultat obtenu est la position $p_{h_{px}}$ en pixels par rapport au bord supérieur gauche de l'écran. Pour avoir la position par rapport au centre de l'écran :

$$p'_{h_{px}} = p_{h_{px}} - \frac{R_h}{2}$$

La position en centimètres est donnée grâce au ratio r du nombre de pixels par centimètre à l'écran :

$$p_{cm} = \frac{p'_{h_{px}}}{r}$$

Enfin, l'angle d'azimut β de la source visuelle perçu par l'utilisateur dépendra non seulement de cette position, mais aussi de la distance d de l'utilisateur au centre de l'écran du téléphone. Pour l'obtenir directement en degrés :

$$\beta = \frac{180}{\pi} * atan(\frac{p_{cm}}{d})$$

Position du visuel

La position du stimulus visuel est conditionnée par celle de l'écran. Nous adoptons l'hypothèse raisonnable qu'un utilisateur, lorsqu'il consomme du contenu audiovisuel, tient le téléphone en face de lui, c'est-à-dire dans le plan médian, à l'azimut 0 °. Quant à l'élévation, nous avons déjà évoqué le fait qu'elle était variable. Néanmoins, pour l'expérience nous nous appuyons sur la publication de [S. Lee, Kang et Shin, 2015] qui nous dit que la flexion moyenne de la tête lors de l'utilisation d'un smartphone est de 39 °. Notre téléphone sera donc placé à 0 ° en azimut, dans le plan médian du sujet, et à -39 ° en élévation par rapport à ses yeux.

Par ailleurs, dans la scène virtuelle, les stimuli visuels sont positionnés à 12 ° d'azimut par rapport à la caméra (l'angle α dans la Figure 3.3), induisant un décalage en azimut par rapport au plan médian du sujet (l'angle β dans la Figure 3.3) qui variera selon sa distance à l'écran (la distance du sujet au téléphone est laissée à sa libre appréciation, voir la Section 3.2.3 pour plus de détails). En pratique dans notre expérience, l'angle β est compris entre 1,17 ° et 2 ° pour une distance au téléphone allant respectivement de 39 à 67cm. Le même décalage est opéré symétriquement dans l'hemi-espace gauche, à -12 ° dans la scène virtuelle, résultant par rapport au plan médian du sujet en un angle en azimut compris entre -1.17 ° et -2 ° sur l'écran.

En résumé, le stimulus visuel sera placé à une position entre 1.17 ° et 2 ° en azimut (et symétriquement entre -1.17 ° et -2 °), et à une élévation de -39 ° par rapport au plan médian du sujet.

Enfin, il faut préciser que le stimulus de l'hélicoptère est orienté de façon à toujours avoir l'avant tourné vers le centre de l'écran (peu importe la trajectoire sonore correspondante). La question ne se pose pas pour le halo, qui est sphérique.

Trajectoires binaurales dans le plan du téléphone

Les trajectoires sonores sont générées à l'aide d'un plugin Unity développé en interne, qui permet de faire de la synthèse binaurale en temps réel à partir d'un fichier monophonique et d'un jeu de HRTF donné. Pour chaque sujet, le volume sonore des trajectoires est calibré de sorte qu'un son statique placé en élévation 0 ° et azimut 0 ° soit perçu à 65dB SPL.

Une première série de trajectoires se situe dans le plan horizontal du centre du téléphone, à -39 ° d'élévation par rapport au regard du sujet, avec une amplitude angulaire en azimut de 30 °. Les trajectoires vont par paire, une qui démarre face au sujet et va vers sa périphérie, et la trajectoire inverse, de la périphérie vers le centre. La détermination de chaque PSSA nous permet ainsi de délimiter la fenêtre d'intégration. Pour s'affranchir d'un éventuel effet d'accoutumance du sujet aux trajectoires, les deux ne sont pas parfaitement symétriques. Celle qui part du centre vers la périphérie va de -5 ° à 25 °, et celle qui va de la périphérie vers le centre va de 30 ° à 0 ° (et la paire de trajectoires dans l'hemi-espace gauche, de 5 ° à -25 ° et de -30 ° à 0 °). La vitesse des trajectoires étant de 2.5 ° par seconde, chaque trajectoire dure 12s.

Trajectoires binaurales en élévation

Nous souhaitons aussi savoir si la décorrélation spatiale entre son et image en élévation, due à un changement de tenue du téléphone par le sujet, influence la fenêtre d'intégration. D'après l'état de l'art qui nous indique une forte tolérance de la perception à des décalages importants en élévation [Thurlow et Jack, 1973 ; Jack et Thurlow, 1973 ; Hendrickx et al., 2015], nous souhaitons essentiellement contrôler non seulement qu'un effet ventriloque a bien lieu en élévation (la tâche d'alignement est toujours faisable par les sujets), mais aussi que l'effet ventriloque en azimut n'en est pas affecté (fenêtre d'intégration identique pour les deux élévations). Nous ajoutons 4 paires de trajectoires, situées au dessus du sujet, à une élévation de 70 °. Cette valeur est choisie expérimentalement et nous paraît être un bon compromis pour offrir à la fois une sensation de hauteur et une sensation de frontalité. Ces trajectoires partagent les mêmes propriétés et variations en azimut que celles dans le plan du téléphone (même amplitude angulaire, même vitesse, mêmes angles de départ et d'arrivée).

Nous sommes par ailleurs confrontés à un problème technique associé au choix de 2 élévations différentes des trajectoires. En effet, chaque trajectoire est générée à distance fixe du sujet, sur une sphère centrée sur sa tête. Cela vient du fait que les HRTF sont mesurées pour chaque direction de l'espace à une distance fixe de l'utilisateur. Or, deux trajectoires horizontales à niveau d'élévation -39 ° et 70 °, si elles parcourent le même angle azimutal, ne parcourent pas du tout la même distance (voir Figure 3.4). En résulte une sensation de « sur-place » pour la trajectoire à 70 °, rendant très difficile la tâche d'alignement demandée. Pour résoudre ce problème, nous générons des trajectoires légèrement différentes. Pour chaque élévation, nous considérons une sphère qui partage son centre avec celui de la section plane à la même élévation sur l'ancienne sphère (voir Figure 3.5). De cette façon, toutes les trajectoires parcourent la même distance. Cette manière de procéder implique que chaque trajectoire, bien que toujours placée à l'azimut et à l'élévation souhaitées, varie légèrement de distance à l'utilisateur durant son parcours, mais nous considérons cette variation comme négligeable.

Pour s'assurer que les deux élévations sont bien différenciées indépendamment de tout stimulus visuel, nous avons procédé à un test informel. Vingt trajectoires sonores, sans visuel, choisies aléatoirement parmi les 8 possibles (2 directions, 2 élévations, 2 hemi-espaces) sont présentées à une dizaine de sujets, qui doivent à chaque fois dire si la trajectoire entendue est au-dessus de la tête ou non. Sur les 10 sujets, 9 d'entre eux répondent correctement pour la totalité des trajectoires. Le dernier sujet fait seulement 3 erreurs, et de surcroît parmi les premières trajectoires qui lui sont présentées. Nous considérons donc que la différence d'élévation entre les trajectoires sonores est effectivement perçue.

En résumé, pour cette expérience du PSSA nous avons 8 trajectoires binaurales, dont chacune est présentée huit fois à chaque sujet, donc 64 trajectoires par type de stimulus (bruit blanc ou bruit d'hélicoptère). Au total donc, 128 trajectoires sont à évaluer pour chaque sujet.

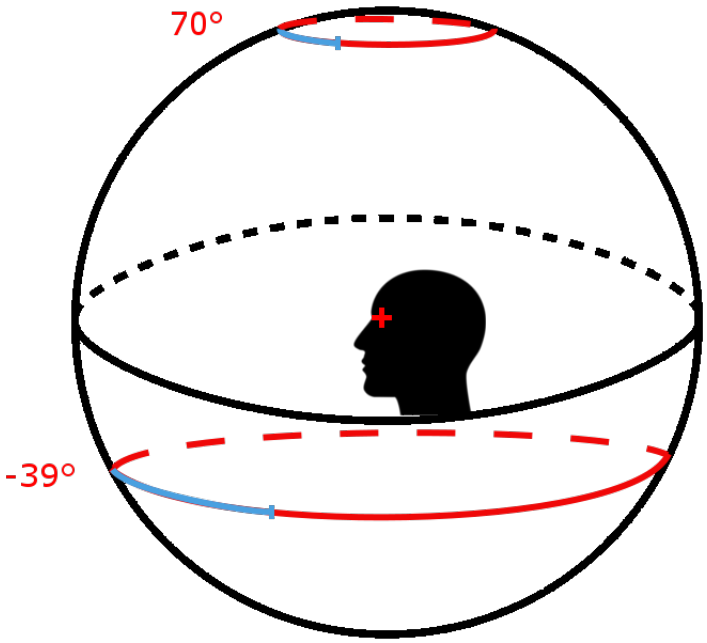


FIGURE 3.4 – Deux trajectoires horizontales, l'une générée à élévation 70 ° , l'autre à élévation -39 ° , qui parcourent le même angle en azimut, ne parcourent pas la même distance.

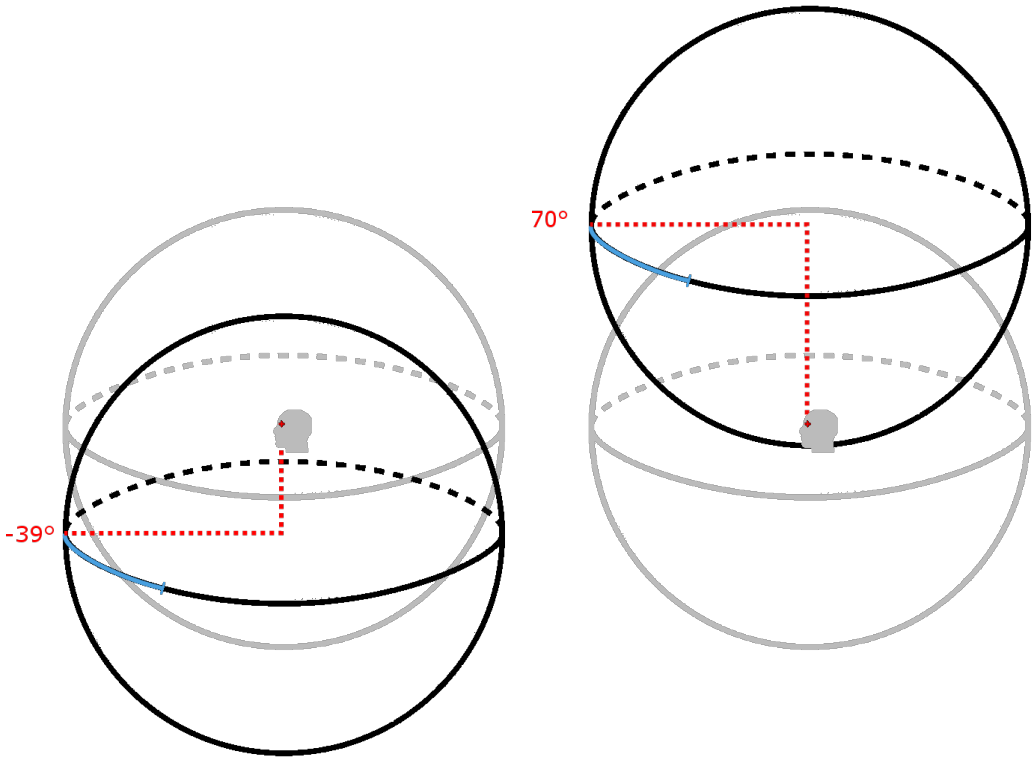


FIGURE 3.5 – Pour chaque élévation, nous générons les trajectoires à partir d'une nouvelle sphère. Ainsi toutes les trajectoires parcourent la même distance.

3.2.3 Dispositif

Le dispositif a été déployé en deux endroits, dans une salle d'expérimentation d'Orange Lannion, et reproduit à l'identique dans une salle d'expérimentation d'Orange Rennes (voir Figure 3.6).

Le téléphone qui sert à l'expérience est posé sur un support inclinable, lui-même posé sur une table. Le sujet est assis sur une chaise à hauteur ajustable face à la table. Il a pour consigne de ne pas toucher au téléphone et de n'interagir avec lui que par l'intermédiaire du contrôleur bluetooth, présenté en section 3.2.1. Malgré son manque de vraisemblance, le choix de ne pas laisser le sujet toucher le téléphone s'impose, afin que le téléphone ne soit pas déplacé de façon intempestive pendant l'expérience, déréglant ainsi le positionnement relatif des stimuli visuels et auditifs.

Par ailleurs, nous avons constaté que dans ce cas, il était beaucoup plus confortable pour le sujet d'ajuster lui-même sa distance au téléphone. Afin de s'assurer que le regard du sujet posé sur le téléphone forme bien un angle de -39 ° avec l'horizontale, nous avons tendu un fil, fixé derrière le sujet (sur le mur pour l'expérience à Lannion, sur une barre verticale pour l'expérience à Rennes) et allant jusqu'à la table devant. Ainsi, après que le sujet a ajusté sa distance, nous réglons la hauteur du siège pour que le fil soit à hauteur des yeux.

Luminosité et volume sonore du téléphone sont d'avance réglés au maximum. Le volume sonore des trajectoires est calibré dans Unity par rapport à ce volume du téléphone. La luminosité générale de la pièce est calibrée à 20 lx (valeur recommandée par l'Union Internationale des Télécommunications pour des expériences perceptives avec contenu audiovisuel [ITU-T, 1998]). La pièce est suffisamment insonorisée pour qu'aucun son extérieur ne soit perçu.

3.2.4 Sujets et déroulement de l'expérience

Trente-quatre sujets participent à l'expérience (de 18 à 58 ans, âge moyen 39,4 ans, 12 femmes). La totalité des participants possèdent une audition normale et une vue normale ou corrigée, à l'exception d'un sujet atteint d'amblyopie unilatérale. Ses résultats étant parfaitement comparables aux autres, nous décidons de conserver ses données. Dix-sept des participants possèdent leurs propres HRTF (mesurées lors du projet BiLi [Rugeles Ospina, Emerit et Daniel, 2015]). Pour les 17 autres, les HRTF d'une tête artificielle Neumann KU 100 sont utilisées. Ceci nous permettra de mesurer l'influence de l'individualisation sur la fenêtre d'intégration. 20 sujets sont des volontaires bénévoles du laboratoire ; les 14 restants sont recrutés via une base de données de volontaires d'Orange, et défrayés par un bon d'achat de 20€.

L'expérience dure au total environ 45 minutes. À son arrivée, le sujet est invité à lire les instructions relatives à l'expérience (disponible en Annexe A.1). Il est informé du but de l'expérience, de la tâche à accomplir, et en particulier des types de trajectoires sonores qu'il aura à écouter (directions –centre vers la périphérie, ou périphérie vers le centre–, élévations –au dessus de sa tête ou au niveau du téléphone–). Quelques

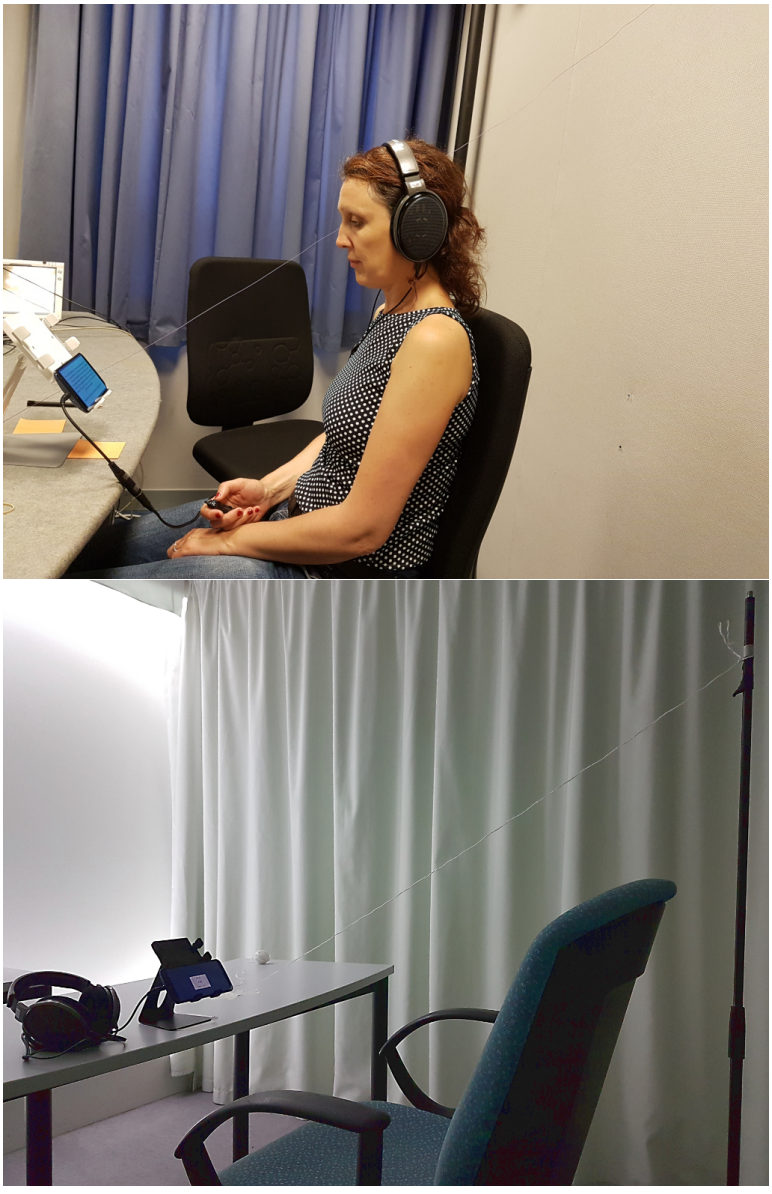


FIGURE 3.6 – Dispositifs expérimentaux, en haut celui de Lannion, en bas celui de Rennes.

questions générales lui sont ensuite posées : âge, préférence manuelle (droitier ou gaucher), éventuels problèmes d'audition ou de vision. Puis il est installé face au dispositif, d'où il règle sa distance, sa hauteur, et enfile le casque audio. Le sujet a pour instruction de garder le contrôleur bluetooth dans la main droite s'il est droitier, gauche s'il est gaucher.

Le sujet se voit présenter deux sessions successives de stimuli, l'une consacrée à l'hélicoptère, l'autre au flash et au bruit blanc. L'ordre des sessions est déterminé de manière à ce que la moitié des sujets commence par l'hélicoptère, l'autre par le bruit blanc. Au sein de chaque session, les trajectoires sont présentées dans un ordre aléatoire.

Pour chaque trajectoire, un bip sonore avertit le sujet de son départ, correspondant également à l'apparition du visuel. Le sujet a pour consigne de fixer des yeux le stimulus visuel (axe du rotor de l'hélicoptère ou centre du halo) et d'indiquer via le contrôleur Bluetooth l'instant dès lequel il perçoit celui-ci aligné (i.e., sur le même axe vertical), voire confondu avec le stimulus auditif. Dès cet instant, le stimulus visuel disparaît et le stimulus auditif s'interrompt. La trajectoire suivante démarre automatiquement, dans un temps aléatoire compris entre 1 et 2 secondes. Si le sujet ne formule aucune réponse pendant une trajectoire, un message s'affiche à l'écran, attendant une validation du sujet pour passer à la session suivante. Une pause est proposée toutes les 8 trajectoires. Néanmoins, à l'exception d'une personne, aucun sujet ne l'a utilisée.

Pour se familiariser avec le protocole, le sujet commence chaque session par une séquence d'entraînement de 8 trajectoires choisies au hasard. Les résultats obtenus pendant l'entraînement ne sont pas retenus pour l'analyse des résultats.

Tous les éléments scénarisés (temps de transition entre deux sessions, déclenchement des bip sonores et trajectoires, mélange aléatoire des stimuli, temps de pause, etc.) sont développés en C# sous forme de scripts dans Unity. L'interface utilisateur est également développée à l'aide des outils intégrés par Unity.

Après les deux sessions, le sujet se voit poser quelques questions relatives au déroulement de l'expérience (disponibles en Annexe A.2). Elles ont pour but de servir à l'appréciation qualitative des résultats.

3.2.5 Hypothèses

Compte tenu des différents résultats exposés dans l'état de l'art, nous formulons les hypothèses suivantes :

— Premièrement, il existe une fenêtre d'intégration en azimut et cette fenêtre entoure la position du visuel telle qu'elle est observée par le sujet sur l'écran.

— Deuxièmement, cette fenêtre d'intégration est la même pour les sujets avec et sans HRTF individualisées.

— Troisièmement, un écart d'élévation entre stimuli visuels et auditifs n'empêche pas la fusion en azimut, du moins dans certaines limites qui restent à quantifier plus précisément.

— Quatrièmement, la cohérence sémantique renforce le lien d'association entre son et image, ce qui implique que la fenêtre d'intégration en azimut est plus large pour des stimuli cohérents sémantiquement.

3.3 Résultats

Des 34 sujets, à raison de 128 trajectoires par sujet, nous collectons 4352 PSSA sous la forme d'un angle en azimut par rapport au centre du sujet (l'angle β, en se référant à la Figure 3.3). Sur ceux-ci, 131 sont retirés des données car provenant de trajectoires ratées pour lesquelles les sujets n'ont pas répondu. Toutes les valeurs de l'hemi-espace gauche sont rapportées à l'hemi-espace droit, afin de pouvoir analyser l'effet de l'hemi-espace sur le PSSA.

Nous supprimons ensuite les valeurs qui sont considérées comme aberrantes (c'est-à-dire qui sont en dehors de l'intervalle de confiance à 95% sur l'ensemble des données). Nous remarquons cependant deux sujets qui ont respectivement 45% et 35% de valeurs aberrantes. Ces deux sujets sont écartés de l'analyse. Nous nous retrouvons avec 3869 valeurs pour 32 sujets.

La moyenne des PSSA par type de stimulus et par direction est représentée en Figure 3.7. On observe que les trajectoires allant du centre vers la périphérie obtiennent un PSSA en moyenne plus bas que pour celles allant de la périphérie vers le centre. En revanche, il y a peu de différence entre les trajectoires de l'hélicoptère et celles du bruit blanc. Une ANOVA à mesures répétées est effectuée en prenant la valeur du PSSA comme variable dépendante. La direction des trajectoires (centre vers périphérie ou périphérie vers centre), l'hemi-espace (gauche ou droite) et l'élévation (-39 ° ou 70 °) sont des facteurs intra-classes, tandis que le type de HRTF (indivualisées ou non) et la rétribution du sujet (payé ou bénévole) sont intégrés à l'analyse comme des facteurs inter-classes. Nous opérons deux analyses séparées pour les stimuli d'hélicoptères et les bruits blancs présentés dans des sessions différentes.

Les résultats suggèrent qu'aucun des facteurs inter-classes n'a d'effet significatif sur le PSSA : ni le type d'HRTF ($F(1, 10)=0.43$, $p=0.52$ pour les stimuli de bruits blancs, $F(1, 8)=0.75$, $p=0.41$ pour les stimuli d'hélicoptères), ni la rétribution ($F(1,10)=0.12$, $p=0.73$ pour les bruits blancs, $F(1,8)=0.13$, $p=0.72$ pour l'hélicoptère).

En ce qui concerne les facteurs intra-classes, la direction de la trajectoire influence significativement le PSSA ($F(1, 10)=13.81$, $p<0.01$ pour les bruits blancs, $F(1, 8)=17.49$, $p<0.01$ pour l'hélicoptère). Les trajectoires « centre vers périphérie » obtiennent des PSSA significativement plus proches du centre que les trajectoires « périphérie vers centre ». Les PSSA moyens des deux types de trajectoires constituent les bornes de notre fenêtre d'intégration.

L'hemi-espace a également un effet significatif, mais uniquement sur les PSSA des stimuli d'hélicoptères ($F(1, 8)=10.78$, $p<0.05$), et pas pour les PSSA des stimuli de bruits blancs ($F(1,10)=1.43$, $p=0.26$). Quand les stimuli d'hélicoptères sont situés dans l'hemi-espace gauche, la fenêtre d'intégration est décalée d'environ 1.7 ° vers la

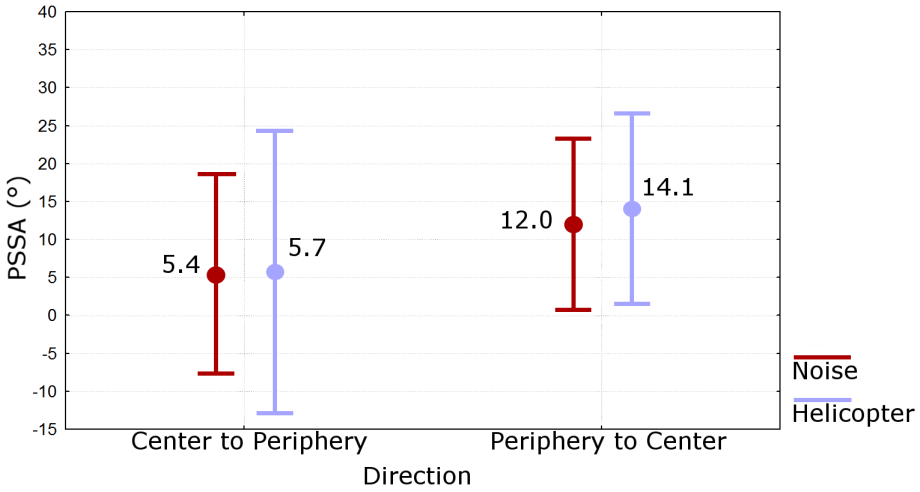


FIGURE 3.7 – PSSA moyen par direction, pour chaque type de stimuli. Les barres verticales sont les intervalles de confiance à 95%.

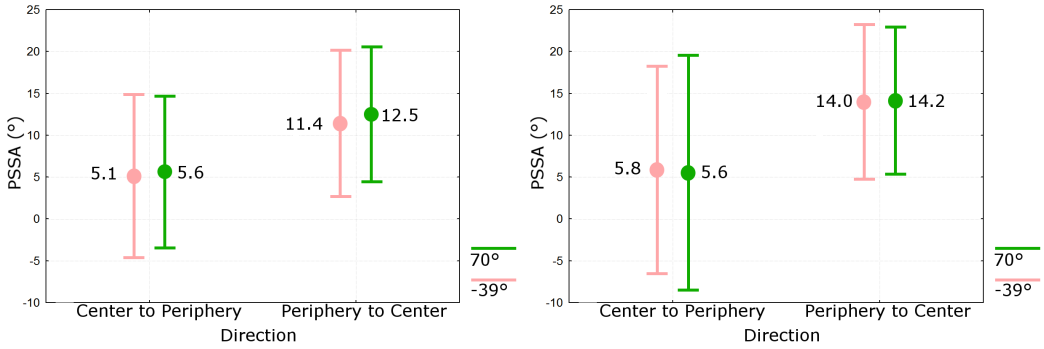


FIGURE 3.8 – PSSA moyen par direction et pour chaque élévation, à gauche pour les stimuli de bruits blancs et flash lumineux, à droite pour les stimuli d'hélicoptères. Les barres verticales sont les intervalles de confiance à 95%.

périphérie comparé à l'hemi-espace droit. En revanche, la taille de la fenêtre ne change pas.

Enfin, l'élévation a un effet significatif, cette fois sur les PSSA des stimuli de bruits blancs ($F(1, 10)=7.50$, $p<0.05$), mais pas d'effet significatif sur les PSSA des stimuli d'hélicoptères ($F(1, 8)=0.11$, $p=0.75$). De plus, il n'y a aucun d'effet croisé entre élévation et direction sur le PSSA, aussi bien pour les stimuli de bruits blancs ($F(1, 10)=0.55$, $p=0.48$) que pour les stimuli d'hélicoptères ($F(1, 8)=2.50$, $p=0.15$). Ces résultats sur l'élévation peuvent être observés sur la Figure 3.8. La fenêtre d'intégration est légèrement, mais significativement décalée vers la périphérie pour les stimuli de bruits blancs élevés.

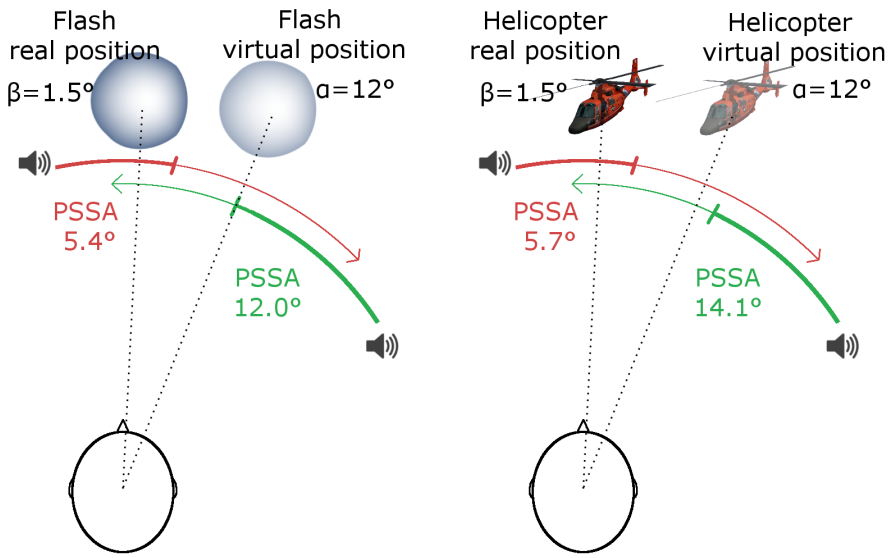


FIGURE 3.9 – Représentation schématique de la fenêtre d'intégration moyenne pour les stimuli de bruits blancs (à gauche) et d'hélicoptères (à droite). Les fenêtre d'intégrations sont décalées en périphérie par rapport au visuel tel qu'il est perçu par le sujet (angle β). Néanmoins, elles encadrent la position du stimulus visuel tel qu'il est placé dans la scène virtuelle (angle α).

3.4 Discussion

3.4.1 Taille et position de la fenêtre d'intégration moyenne

L'effet significatif de la direction des trajectoires sur le PSSA confirme l'existence d'une fenêtre d'intégration, d'une largeur en moyenne de 6.6 ° pour les stimuli de bruit blanc et flash lumineux, et 8.4 ° pour les stimuli d'hélicoptère. Ces valeurs nous semblent être du même ordre de grandeur, quoiqu'un peu plus grandes, que les fenêtres d'intégration trouvées par [NGUYEN, 2012] (qui pour rappel étaient de 2.5 °). Nos résultats sont aussi très proches de ceux de [LEWALD, EHRENSTEIN et GUSKI, 2001], puisque leur fenêtre d'intégration moyenne vaut environ 6 ° (en se souvenant toutefois que leur protocole est légèrement différent, puisque c'est le stimulus visuel qui bouge et le stimulus auditif qui est statique).

Cependant, contrairement à notre première hypothèse, nos fenêtres n'entourent pas le stimulus visuel présenté sur l'écran (voir Figure 3.9). Nous l'avons dit, en fonction de la distance du sujet au téléphone, le visuel est positionné sur l'écran à un angle en azimut compris entre 1.5 ° et 2 ° par rapport au plan médian du sujet. En moyenne, les bornes inférieures et supérieures des fenêtres sont respectivement de 5.4 ° et 12 ° pour le bruit blanc, et 5.7 ° et 14.1 ° pour l'hélicoptère. Les fenêtres sont donc décalées vers la périphérie. [NGUYEN, 2012] avait également un décalage de ses fenêtres par rapport au visuel, mais vers le centre (dans les conditions expérimentales similaires aux nôtres). Dans notre cas, nous constatons que ce décalage coïncide avec la position du stimulus visuel dans la scène virtuelle. Autrement dit, en reprenant les notations de la Figure 3.3, la fenêtre d'intégration encadre α au lieu de β.

Dans son manuscrit de thèse sur la perception et l'attention au cinéma, Claude Bailblé évoque un mécanisme d'attention permettant d'interpréter le décalage obtenu dans nos résultats [Bailblé, 1999] :

> « Lorsque la lumière s'éteint, l'espace de la salle est comme scotomisé : la surface lumineuse de l'écran attire les regards et empêche toute errance de l'œil. La salle est alors rejetée dans l'oubli, seul l'espace scénique est investi [...]
> En multiphonie, à chaque fois que la source sonore ne s'ajuste plus à l'image, le spectateur reprend conscience de l'espace physique tridimensionnel de la salle et éprouve une gêne. L'espace de la salle ne devrait pas faire partie de l'espace scénique, aussi le va-et-vient entre ces deux espaces perturbe-t-il la plongée imaginaire dans la fiction. »

Selon Baiblé, l'espace réel qui entoure le sujet, considéré comme gênant, serait comme évacué par le sujet (« scotomisé »), qui ne considèrerait alors plus que l'espace virtuel. D'un point de vue spatial, le spectateur se projetterait alors à la place de la caméra. Si cette théorie s'est construite sur l'analyse de l'attention et la perception au cinéma, nous pensons qu'elle peut être valable dans notre cas, et qu'elle peut être à l'origine du décalage de la fenêtre d'intégration. Ce résultat est particulièrement intéressant dans la mesure où il est déjà une marque d'immersion du sujet. Si nous n'avons aucune indication sur le fait que le binaural soit à l'origine de ce qui paraît être une immersion dans la scène (le sujet pourrait très bien percevoir le visuel à l'angle α indépendamment de tout stimulus sonore), nous savons au moins que sa présence ne la détériore pas. Une expérience purement visuelle, où l'on demanderait au sujet de reporter la position à laquelle il perçoit un objet dans une scène virtuelle, nous permettrait d'en apprendre plus sur cette question.

Quoi qu'il en soit, le fait que le sujet considère la scène virtuelle du point de vue de la caméra serait un atout dans la construction d'une scène audiovisuelle à destination du grand public : il ne serait pas nécessaire de se préoccuper de la position relative de l'utilisateur par rapport au téléphone, il suffirait simplement de considérer le point d'écoute de la caméra.

Cette conclusion va dans le sens de plusieurs conversations informelles tenues au cours de la thèse avec des professionnels de l'audiovisuel, du jeu vidéo et du son binaural. Ceux-ci doivent déjà considérer concrètement la question du point d'écoute lorsqu'ils élaborent des contenus audiovisuels avec du son binaural. Dans l'état actuel des choses, ils adoptent tous tacitement cette règle : le point d'écoute est placé à l'endroit de la caméra. Nos résultats corroborent et appuient cette pratique déjà en vogue.

3.4.2 Effet de l'individualisation des HRTF

Un autre résultat important de cette expérience est l'absence d'effet significatif de l'individualisation des HRTF sur le PSSA. Les sujets avec des HRTF individualisées ont une perception spatiale relative du son et de l'image comparable à celle des sujets avec des HRTF standard. Là aussi, le résultat est important pour l'utilisation de la technologie binaurale dans des applications audiovisuelles destinées au grand public, puisqu'il indiquerait que l'absence d'individualisation n'impacte pas l'intégration spatiale auditivo-visuelle.

3.4.3 Effet de l'hemi-espace

L'influence significative de l'hemi-espace sur le PSSA pour les stimuli d'hélicoptère va à l'encontre de tous les résultats trouvés dans la littérature. Plusieurs expériences vont même jusqu'à négliger un des deux hemi-espaces, en partant du principe que l'effet ventriloque est symétrique [Komiyama, 1989 ; Lewald, Ehrenstein et Guski, 2001 ; Hendrickx et al., 2015 ; André et al., 2014]. D'autres publications comme [Jackson, 1953 ; Jack et Thurlow, 1973] montrent un effet de l'hemi-espace, mais toujours attribué à des réverbérations de la salle, et non symétrique. Dans notre cas cette interprétation n'est pas valable, puisque les stimuli sonores sont rendus sur casque. Nous n'avons pas d'explication solide qui pourrait étayer cet effet significatif. Certaines études dans le domaine de la communication suggèrent que la perception d'un stimulus en mouvement peut être affectée selon que sa direction soit concordante ou non avec le sens de lecture (voir par exemple [E. Li et Briley, 2011]). Ici cependant les stimuli en mouvement sont uniquement auditifs. Une étude complémentaire nécessite d'être menée pour conclure sur ce résultat inattendu.

3.4.4 Effet de l'élévation

L'influence de l'élévation sur le PSSA obtenu avec les stimuli de bruits blancs révèle que la fenêtre d'intégration pour les trajectoires élevées est décalée en périphérie par rapport à celle des trajectoires dans le plan du téléphone. Pour autant, si ce décalage est significatif d'un point de vue statistique, il n'est numériquement que de 0.8°, en dessous des seuils de perception auditive angulaire habituellement considérés aux abords du plan médian [Woodworth et Schlosberg, 1954]. Pour cette raison, il nous est difficile d'interpréter ce résultat. Là encore, une étude plus poussée est nécessaire.

Quoi qu'il en soit, cette similitude de taille entre les fenêtres d'intégration indique un effet ventriloque en azimut aussi efficace quelle que soit l'élévation. Ce résultat est cohérent avec celui trouvé notamment par [Jack et Thurlow, 1973], et cité dans le chapitre 2, qui présentaient un visuel devant le sujet avec une source sonore placée derrière (élévation 180°) et des angles en azimut variables. Ils montraient que la perception de cohérence entre son et image décroît à mesure que le stimulus sonore s'écarte en azimut du plan médian, dans les mêmes proportions qu'un stimulus sonore s'écartant en azimut de la position frontale.

Pour résumer, ces résultats suggèrent qu'une différence de position verticale du téléphone dans les mains de l'utilisateur ne devrait pas dégrader la cohérence spatiale perçue d'un objet audiovisuel. Toutefois, dans l'optique de se rapprocher davantage d'une utilisation réelle de smartphone, il serait intéressant de tester la perception du sujet dans le cas où l'élévation relative entre le son et l'image change au cours de la trajectoire sonore, pour simuler un changement de position du téléphone au cours d'une même utilisation.

3.5 Conclusion : les réponses utiles à la suite de la thèse

Un des objectifs de cette thèse est d'évaluer l'apport du binaural sur smartphone dans un contexte proche d'une utilisation grand public. Les résultats obtenus avec cette première expérimentation y semblent favorables : absence de dégradation de la fenêtre d'intégration auditivo-visuelle par l'utilisation de HRTF non individuelles, possibilité de considérer le point d'écoute de l'utilisateur comme s'il était plongé directement dans la scène virtuelle (sans tenir compte de sa distance à l'écran dans l'environnement réel), et possibilité enfin de s'affranchir de la position verticale incertaine du téléphone dans les mains de son utilisateur.

Dans les prochains chapitres, nous abordons le point central de notre problématique. Les résultats obtenus seront utilisés lors de la conception de l'application mobile qui servira de support à nos expérimentations.

Deuxième partie

Mesure de l'apport du binaural sur mobile

Chapitre 4

Qualité sonore, qualité d'expérience et contexte

4.1 Introduction

La plupart des tests d'évaluation du binaural que nous avons vus dans le chapitre 2 se basent sur les performances de localisation. Mais quelle est la pertinence de la tâche de localisation au regard de l'utilisation du binaural dans une expérience audiovisuelle telle que le visionnage d'un film, la participation à une visioconférence ou une partie de jeu vidéo ? Il est peu courant dans ce type d'applications de se voir attribuer une tâche explicite de localisation de sources sonores ou audiovisuelles. Par ailleurs, quand bien même cette tâche surviendrait, par exemple dans un jeu vidéo, les performances, conditionnées ou non par la présence du binaural, ne reflèteraient pas nécessairement un ressenti positif de l'expérience vécue par l'utilisateur. Il apparaît dès lors que la notion subjective de ressenti est de première importance pour évaluer l'apport du binaural dans une application mobile. Plusieurs domaines de recherche s'accordent pour désigner le niveau d'appréciation ressenti par l'utilisateur par le terme de « qualité ». D'une certaine façon, elle est un moyen de quantifier la distance qui sépare un concepteur, dans la façon dont il prévoit les fonctionnalités de son application, de l'utilisateur, dans la façon dont il les utilise [De Moor, 2012].

Pour commencer nous définissons dans la section 4.2 la qualité dans son acception générale. Dans la section 4.3, nous nous penchons sur la qualité sonore, ses définitions, ses attributs, ses méthodes d'évaluation et les expériences ayant déjà mesuré l'apport du binaural par rapport à d'autres systèmes de sonorisation. Dans la section 4.4, nous présentons la notion plus vaste de qualité d'expérience, concept utilisé pour définir et quantifier l'appréciation globale d'une expérience par un utilisateur. Cette section sera l'occasion de concevoir le binaural comme un élément parmi d'autres intégré au sein d'une expérience d'application audiovisuelle. Enfin, la qualité d'expérience étant fortement sujette à variations selon le contexte d'utilisation, nous passons en revue dans la section 4.5 la définition du contexte et des facteurs d'influence de la qualité d'expérience. Nous nous focaliserons en particulier sur l'influence de ces facteurs dans des contextes mobiles. Enfin, la question de la validité écologique des données liée à la méthode de déploiement est évoquée, pour laquelle on envisage généralement de replacer l'expérience dans un contexte d'utilisation réaliste.

4.2 La qualité

4.2.1 Définitions de la qualité

La définition de la qualité, indépendamment de son rattachement au son, a été discutée par [S. Möller et Raake, 2014]. Ils observent que le terme a radicalement évolué depuis le début des années 2000 [H. Martens et M. Martens, 2001]. Avant, elle était par exemple définie par [ISO, 1994] comme « la totalité des caractéristiques d'une entité qui portent sur ses capacités à satisfaire des besoins déclarés ou implicites » [1]. Cette « totalité des caractéristiques » ferait référence à l'origine latine « qualitas », qu'on traduirait aujourd'hui par « caractère » [Blauert et Jekosch, 2003] [2]. La définition a été ensuite modifiée en 2000 par l'ISO pour : « aptitude d'un ensemble de caractéristiques [...] intrinsèques à satisfaire des exigences », où une caractéristique est ici un « trait distinctif », et le terme « intrinsèque », opposé au terme « attribué », signifie « présent dans quelque chose, notamment en tant que caractéristique permanente » [3] [ISO, 2000]. La qualité, qui décrivait au départ un ensemble de caractéristiques et ne dépendait en cela que de l'entité en question, se rapproche désormais de l'utilisateur et devient un niveau de satisfaction par rapport à des exigences exprimées. Depuis, la définition de l'ISO n'a pour ainsi dire pas changé dans la dernière occurrence de la norme en 2015 : « l'aptitude d'un ensemble de caractéristiques [...] intrinsèques à un objet [...] à satisfaire des exigences » [4], l'objet pouvant être « n'importe quoi qui soit perceptible ou concevable. », « matériel [...], immatériel [...] ou imaginaire » [5] [ISO, 2015].

Dans [Jekosch, 2006, p.15], une autre définition est donnée (également adoptée par [Raake et Egger, 2014]) : « le résultat d'un jugement sur la composition d'une entité telle qu'elle est perçue, en comparaison de la composition qui était désirée » [6]. Ici la qualité se rapproche encore davantage de l'utilisateur, puisqu'elle dépend désormais de ses attentes, c'est-à-dire de la façon dont le besoin exprimé aurait pu, dans son imagination, être assouvi grâce à l'entité. L'introduction de la perception est également très importante, car elle exprime le fait que la qualité peut varier selon les conditions dans lesquelles l'entité est présentée. La qualité devient donc relative à un point de vue.

Dans [Le Callet et al., 2013], une définition de la qualité est formulée par le Réseau Européen sur la Qualité d'Expérience dans les systèmes et services multimédias : « l'aboutissement d'un processus de comparaison et de jugement chez un individu. Cela inclut la perception, la réflexion sur cette perception et la description de ce qui en découle. À l'opposé des définitions qui voient la qualité comme "qualitas", i.e., comme un ensemble de caractéristiques intrinsèques, nous considérons la qualité en termes de valeur ou d'excellence évaluée, de degré d'accomplissement d'un besoin, et en termes

1. *totality of characteristics of an entity that bear on its ability to satisfy stated and implied needs.*
2. acception qu'on peut d'ailleurs retrouver dans le mot qualité en français
3. traduction française officielle
4. *degree to which a set of inherent characteristics [...] of an object [...] fulfils requirements*
5. *anything perceivable or conceivable [...] material [...], immaterial [...] or imagined*
6. *results from the "judgment of the perceived composition of an entity with respect to its desired composition".*

d'"événement qualité" » [7]. Cette définition, dans la continuité de la précédente, appuie l'aspect éphémère de la qualité, avec notamment cette notion d'événement qualité, qui replace le jugement dans un endroit et à un moment précis.

4.2.2 Processus de formation du jugement de qualité

N'importe quel utilisateur est capable de formuler un jugement de qualité à propos d'une entité qui lui est présentée. Mais quel est le processus interne qui lui permet de le former ? [Le Callet et al., 2013] en proposent une conceptualisation. Ils identifient deux chemins par lesquels le processus se construit : le chemin de perception et le chemin de référence. La Figure 4.1 présente un schéma de ce processus complexe. Le chemin de perception considère le passage d'une entité depuis le monde physique jusqu'au domaine de la perception, par le biais des signaux physiques émis par l'entité qui atteignent les organes sensoriels de l'utilisateur. En termes de qualité, l'entité est d'abord décrite dans le monde physique comme un ensemble d'« éléments de qualité », ou qualitas si l'on reprend le terme de la section précédente, et dans le domaine de la perception comme un ensemble de « caractéristiques de qualité » (voir [Raake et Egger, 2014] pour les définitions détaillées de ces deux termes). Le signal physique est traité via des processus bas niveau de la perception, pour devenir un percept soumis aux contraintes du chemin de référence. Celui-ci reflète la nature temporelle et contextuelle du processus de formation du jugement, qui permet de comparer le percept à une référence interne, dépendante de la mémoire des expériences précédentes de l'utilisateur. En d'autres termes, la qualité perçue est ici comparée à la qualité désirée. À l'issue de ces deux chemins, l'entité, décrite intérieurement de façon complète, éventuellement quantifiée sous plusieurs aspects, peut susciter un jugement de qualité. Pour un approfondissement sur la formation du jugement de qualité, notamment d'un point de vue cognitif, voir [Jekosch, 2006 ; Raake, 2007 ; Raake et Egger, 2014]).

4.3 La qualité sonore

4.3.1 Définitions

Lorsque l'entité perçue par l'utilisateur est un événement sonore, on parle de qualité sonore [Raake, 2016]. La norme [ITU-R, 2015b] parle aussi de « qualité audio de base » pour les systèmes mono, stéréo et multicanal, qui est définie comme un attribut unique et global utilisé pour juger de toute différence perçue entre deux objets sonores.

D'une façon plus élaborée, [Letowski, 1989] nous donne la définition suivante : « La qualité sonore est cette évaluation d'une image auditive exprimée par l'auditeur en termes de satisfaction ou d'insatisfaction. La qualité sonore peut être jugée en comparant des images produites par plusieurs stimuli externes, ou en référant l'image perçue

7. *the outcome of an individual's comparison and judgment process. It includes perception, reflection about the perception, and the description of the outcome. In contrast to definitions which see quality as "qualitas", i.e., a set of inherent characteristics, we consider quality in terms of the evaluated excellence or goodness, of the degree of need fulfillment, and in terms of a "quality event".*

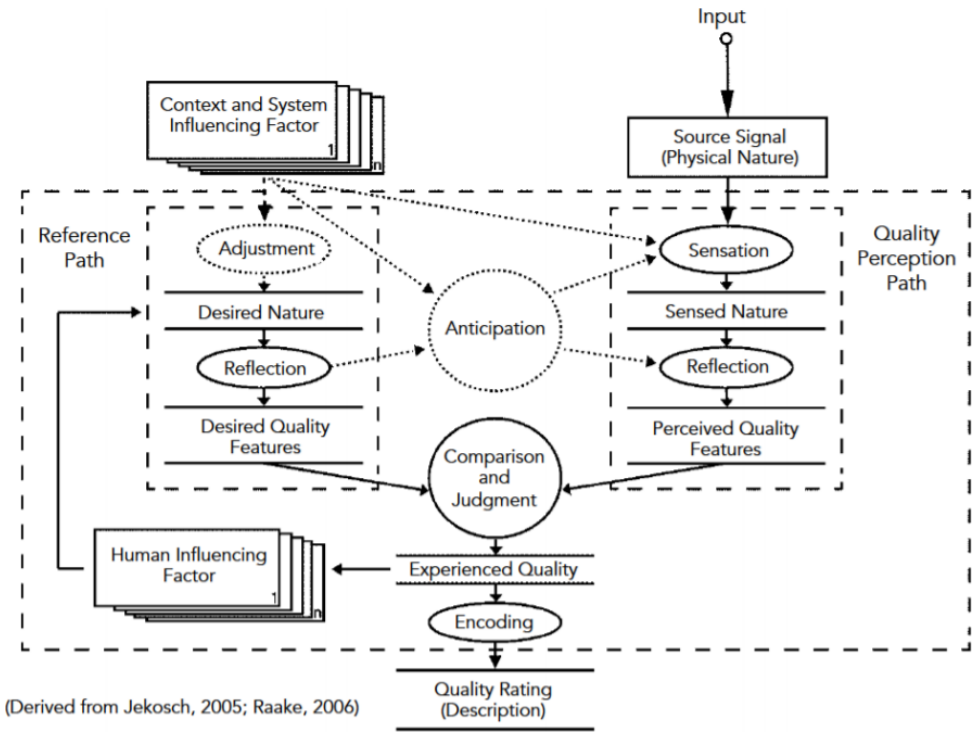


FIGURE 4.1 – Schéma du processus de formation de la qualité d'après [LE CALLET et al., 2013]. À droite le chemin de perception, et à gauche le chemin de référence, qui mènent tous deux à la comparaison et au jugement, pour aboutir à une évaluation de la qualité.

au concept résidant dans la mémoire de l'auditeur »[8]. Cette définition correspond à celles sur la qualité : un jugement subjectif qui peut reposer sur une référence interne de l'auditeur.

Tout comme les entités de la section précédente, un son peut s'envisager par le prisme du monde physique ou de la perception (voir par exemple [LETOWSKI, 1989 ; W. MARTENS, 2001 ; BLAUERT et JEKOSCH, 2003 ; SPORS et al., 2013]). Là aussi, la qualité sonore ne porte pas la même définition selon qu'elle soit considérée de l'un ou l'autre point de vue. Comme dans la section précédente, on a d'un côté l'élément de qualité, propre à l'entité sonore, indépendante de tout jugement de valeur, et de l'autre la caractéristique perceptive de qualité, ancrée dans l'espace et dans le temps par un avis de l'utilisateur.

Pour aller plus loin, et dans la continuité du modèle de formation du jugement de qualité présenté dans la section précédente, [BLAUERT et JEKOSCH, 2012] répartissent différents attributs de la qualité sonore selon le niveau perceptif auquel ils sont sollicités :

— la qualité auditive ;

— la qualité de la scène auditive ;

8. *Sound quality is that assessment of auditory image in terms of which the listener can express satisfaction or dissatisfaction with that image. Sound quality can be judged by comparing images produced by several external stimuli or by referencing a perceived image to the concept residing in the listener's memory.*

— la qualité acoustique ;
— la qualité de communication auditive.

Le tableau 4.1 résume le contenu de chacune d'entre elles. La qualité auditive émane du plus bas niveau d'abstraction de l'individu. Elle permet de juger des attributs sonores tels que l'intensité, la hauteur, la rugosité d'un son, son timbre, etc., indépendamment de toute interprétation cognitive. Cette première couche fait état d'une écoute discrétisante (ou analytique, destinée à évaluer seulement certaines composantes sonores, indépendamment de leur appartenance à un objet) plutôt que syncrétique, qui est pourtant le mode d'écoute habituel de la vie quotidienne.

L'écoute syncrétique implique le groupement mental des attributs en objets sonores, eux-mêmes intégrés dans une scène auditive unifiée, et nécessite un niveau d'abstraction supplémentaire. On parle alors de qualité de la scène auditive. À ce stade, on retrouve des effets tels que l'effet de précédence, l'effet cocktail party, des effets d'attention contrôlée, etc. [Blauert et Jekosch, 2012] argumentent que c'est à ce niveau que les ingénieurs du son travaillent : faciliter la localisation et l'identification des sources sonores, l'intelligibilité des voix, favoriser une balance des timbres équilibrée, etc. Les caractéristiques propres à la qualité sonore mobilisées ici sont donc l'impression de spatialité, l'immersion, le sentiment de présence, la plausibilité perceptive, etc.

La qualité acoustique concerne les caractéristiques physiques (i.e., acoustiques) d'un signal et la façon dont elles sont perçues par l'être humain. Ces caractéristiques, pour être reliées à un jugement de la part de l'auditeur, requièrent une abstraction mathématique de haut niveau. La qualité associée aux propriétés acoustiques d'un signal peut servir à décrire par exemple la validité d'un signal à l'issue d'une chaîne de transmission (e.g., un convertisseur analogique numérique, ou même un système purement acoustique comme une salle réverbérante). De nombreux systèmes existent qui infèrent automatiquement une note de qualité acoustique fondée uniquement sur les propriétés physiques du signal. Les caractéristiques propres à ce niveau de qualité sont le niveau de pression acoustique, la réponse impulsionnelle, le temps de réverbération, la fonction de transmission, etc.

Enfin, la qualité de communication auditive mobilise un niveau d'abstraction encore plus important. Elle s'applique au signe, c'est-à-dire à l'unité de sens véhiculée par l'événement auditif. [Blauert et Jekosch, 2012] font état de trois éléments essentiels pour que ce signe soit correctement compris : un événement auditif porteur de sens, un auditeur, et que cet auditeur porte en lui un concept de l'objet qui lui serve de référence cognitive. La réception d'un son demande alors un certain travail d'interprétation, d'analogie, de connotations pour que lui soit attribuée une signification. Ce niveau est de première importance, par exemple pour les sound designers, qui doivent concevoir des sons chargés d'un sens bien précis. La pertinence d'un son à communiquer ce sens fait donc partie des caractéristiques propres à ce niveau de qualité.

4.3.2 Attributs sonores

La proposition de découpage de [Blauert et Jekosch, 2012] nous permet de mettre en évidence le caractère protéiforme d'une entité sonore perçue par un auditeur, et de

Conceptual Aspect	*Example of Issues*	*Suitable Measuring Methods*
Auditive quality *Classical Psychoacoustics*	Perceptual properties such as loudness, roughness, sharpness, pitch, timbre, spaciousness	*Indirect scaling* : thresholds, difference limens, points of subjective equality *Direct scaling* : category scaling, ratio scaling, direct magnitude estimation
Aural-scene Quality *Perceptual Psychology*	Identification and localization of sounds in a mixture, speech intelligibility, audio perspective incl. distance cues, scenic arrangement, tonal balance, aural transparency	*Discretic* : semantic differential, multi-dimensional scaling. *Syncretic* : scaling of preference, suitability, and/or appropriateness, benchmarking against target sounds
Acoustic Quality *Physics*	Sound-pressure level, impulse response, transmissions function, reverberation time, sound-source position, lateral-energy fraction, inter-aural cross correlation	Instrumental measurements with physical equipment for the measurement of elasto-dynamic vibrations and waves, including appropriate signal processing
Aural-communication Quality *Communication Sciences*	Product-sound quality, comprehensibility, usability, content quality, immersion, assignment of meaning, dialogue quality	Psychological (cognitive) tests, particularly in realistic use cases, e.g., the product in use, the audience in concert, etc., questionnaires, dialogue tests, comprehension test, usability tests, market surveys

TABLE 4.1 – Un résumé des catégories de la qualité sonore telles que présentées par [BLAUERT et JEKOSCH, 2012]. Notons que l'immersion est incluse dans la dernière catégorie, alors que le texte de [BLAUERT et JEKOSCH, 2012] y fait explicitement référence comme faisant partie de la deuxième. Il s'agit peut-être d'une erreur, à moins que les auteurs l'incluent dans les deux.

la multiplicité des points de vue possibles pour évaluer sa qualité. [Letowski, 1989] explique l'intérêt d'évaluer des attributs séparément plutôt que tous ensemble : le traitement auditif des événements par l'être humain est limité en capacité et éprouve des difficultés à comparer plusieurs attributs en même temps. Plus encore, les différences entre les attributs les plus sensibles peuvent masquer les différences entre les autres attributs. Il vaut mieux donc limiter les attributs que l'on veut comparer, et maintenir à l'identique ceux dont on ne souhaite aucun jugement.

Dans ce cas, le choix des attributs constitue un problème complexe. [Berg et Rumsey, 1999] avancent que les attributs ne doivent pas seulement décrire correctement les aspects du son qu'on veut mesurer, mais également être compréhensibles pour tous les sujets de la même façon. Ils recensent les différentes méthodes pour déterminer les attributs sonores. Ils considèrent deux catégories principales : les attributs directement fournis aux sujets par l'expérimentateur (attributs « pourvus »), et les attributs élaborés par les sujets eux-mêmes (attributs « extraits »).

Attributs « pourvus »

Un des avantages de l'expérimentateur qui fournit lui-même les attributs est la rapidité de mise en œuvre. Par ailleurs, avec cette façon de procéder, l'expérimentateur peut attirer l'attention du sujet sur des attributs qui ont déjà été éprouvés par le passé (possiblement même des attributs « extraits » lors de précédentes recherches). On en trouve un exemple chez [Toole, 1985], qui reprend les attributs obtenus par [Gabrielsson, 1979], eux-mêmes en partie extraits auprès de sujets, puis filtrés par des experts. Dans [Staffeldt, 1974], les participants doivent comparer des stimuli sonores selon 35 attributs différents, élaborés sur la base de précédents tests d'écoutes, d'interviews et d'expériences pilotes. Autre exemple enfin dans [Berg et Rumsey, 2002], qui reprennent des attributs de leur précédente expérience [Berg et Rumsey, 2001], mélangés à des attributs résultant d'une nouvelle expérience.

Dans cette même logique, s'appuyer sur des experts pour définir les attributs peut s'avérer pertinent. Par exemple dans [Gabrielsson et Sjögren, 1979], 6 expériences sont présentées, où des attributs sonores, préalablement sélectionnés par des experts du domaine de l'audio, sont soumis à des sujets pour émettre un jugement de préférence sur des systèmes de reproduction sonore. Les auteurs utilisent ensuite une analyse en composantes principales pour extraire à partir des résultats les attributs principaux qui ont guidé leurs réponses.

L'inconvénient de fournir des attributs reste bien sûr le biais potentiel introduit par une présélection arbitraire, qui peut orienter le sujet sur des aspects qui ne lui auraient pas semblé pertinents autrement, et le danger potentiel de mettre le sujet face à un attribut qu'il ne comprend pas.

Attributs « extraits »

À l'inverse, les attributs « extraits » ont pour avantage d'être accessibles à la compréhension des sujets et conformes à leur point de vue. L'inconvénient reste une mise en œuvre longue, qui peut s'étirer sur plusieurs semaines pour certaines méthodes. [Berg et Rumsey, 1999] subdivisent ces méthodes en trois sous-catégories : 1) celles pour lesquelles un groupe de sujets se met d'accord sur une liste commune d'attributs ; 2) celles pour lesquelles les sujets élaborent individuellement les attributs ; et enfin 3) celles qui utilisent d'une façon ou d'une autre une analyse multidimensionnelle fondée sur des similarités ou différences entre les stimuli.

Dans la première catégorie, on trouve par exemple [Stone et al., 1974], qui présentent la méthode QDA (*Quantitative Descriptive Analysis*), utilisée initialement dans le domaine alimentaire. Un panel d'individus est entraîné pour sélectionner les attributs qui correspondront le mieux à un produit, sur la base de discussions dirigées par un chef de panel. À l'issue de cette sélection, les sujets notent différents produits au travers de ces attributs, via des échelles de notation. Une série d'analyses (ANOVA, calcul de coefficients de corrélation, analyse en composantes principales) permet enfin de mettre en lumière les attributs principaux et d'éliminer les redondances. Cette méthode de QDA a été reprise dans le domaine de la qualité sonore, e.g., [Mattila, 2001 ; Zacharov et Koivuniemi, 2001b ; Zacharov et Koivuniemi, 2001a ; Bech et G. Martin, 2005 ; Lorho, 2005a ; Francombe, Brookes, Mason et Woodcock, 2016 ; Francombe, Brookes et Mason, 2017].

Dans la seconde catégorie, les sujets élaborent leurs attributs individuellement, sans être influencés par l'intervention des autres sujets. La technique de *Free-Choice Profiling* est une de ces méthodes, et possède l'avantage de ne pas forcer l'accommodement des sujets à un vocabulaire technique, ni de devoir faire appel à des sujets experts. Après l'écoute des stimuli à comparer, le sujet peut exprimer librement les impressions qui lui viennent à l'esprit. À l'issue de l'expérience, l'expérimentateur doit analyser les verbatims récoltés. Il peut le faire manuellement, selon le niveau de complexité des paroles recueillies, comme dans [Guastavino et Katz, 2004], ce qui exige un travail de regroupement et de réduction syntaxiques (rassembler les synonymes, rapporter les phrases sémantiquement similaires à des lemmes communs). Cela constitue sans doute la partie la plus longue, comme en témoignent [Dubois, 2000 ; Guastavino et Chemiliée, 2003]. À l'issue de cette phase, une analyse statistique permet de regrouper les attributs et d'en déterminer les éléments principaux. La même méthode est utilisée par [Lorho, 2005b], dans laquelle les sujets doivent toutefois parvenir eux-mêmes à l'énonciation d'attributs précis, à l'issue d'une phase de discussion de 3 à 4 heures. Les expérimentateurs sélectionnent dans ce cas par des tests préliminaires les sujets les plus aptes à décrire efficacement les stimuli. Dans la même catégorie de méthodes, la technique de *Repertory Grid* est amenée dans le domaine de la qualité sonore par [Kjeldsen, 1998], et utilisée pour la première fois par [Berg et Rumsey, 1999] dans le champ du son spatialisé. Elle propose au sujet de construire son propre lexique, en comparant les stimuli par groupe de trois. À chaque fois, le sujet doit expliciter en quoi deux des stimuli sont différents du troisième. Une fois toutes les combinaisons de stimuli présentées, le sujet doit les noter un par un sur des échelles bipolaires reprenant les termes élaborés. La difficulté de cette méthode, rapportée par [Berg et Rumsey, 1999], concerne l'analyse statistique, puisque chaque sujet possède ses propres termes. Plusieurs solutions sont possibles, comme par exemple une analyse de

corrélation sémantique entre les attributs des différents sujets. En plus de l'expérience de [Berg et Rumsey, 1999], développée plus tard dans [Berg et Rumsey, 2000 ; Berg et Rumsey, 2001], d'autres travaux dans le domaine sonore ont utilisé cette technique, comme [Berg et Rumsey, 2002 ; Berg, 2005 ; Choisel et Wickelmaier, 2006 ; Geier et al., 2010].

Dans les deux catégories précédentes, on supposait que l'auditeur est capable d'évaluer un son sur la base de descripteurs verbaux. Le principe général des méthodes indirectes d'extraction d'attributs consiste à demander au sujet une évaluation non-verbale de stimuli. Les résultats de cette évaluation permettent alors une représentation des données dans un espace multidimensionnel, qui requiert un travail complémentaire d'interprétation de la part des expérimentateurs (accompagné ou non d'analyses statistiques additionnelles, d'une tâche d'énonciation verbale d'attributs par les sujets ou par des experts, etc.) pour comprendre quels sont les principaux attributs qui sous-tendent cette organisation. Dans [Eisler, 1966] par exemple, le sujet doit attribuer une note de qualité globale à plusieurs stimuli, à l'issue de quoi l'expérimentateur applique une analyse factorielle aux données pour mettre en lumière les principaux attributs. Dans [B. J. McDermott, 1969 ; T. Nakayama et al., 1971 ; Gabrielsson, 1979 ; W. L. Martens et Zacharov, 2000], le sujet doit évaluer la similarité entre des stimuli présentés par paires. Une analyse par positionnement multidimensionnel (*multidimensional scaling*) permet alors de représenter les stimuli dans un espace perceptif, organisés selon les similarités obtenues. Plutôt que par paires, les stimuli peuvent aussi être comparés par groupe de trois, comme dans l'analyse dite *Perceptual Structure Analysis* de [Choisel et Wickelmaier, 2006], ou même plus, comme avec la méthode de *Similarity Picking with Permutation of References (SPPR)* de [Michaud et al., 2013], où le sujet doit comparer un sous-ensemble des stimuli accompagnés d'une référence. Enfin, dans [Giacalone et al., 2017], la méthode de *Projective Mapping* consiste à directement répartir les stimuli sur un espace en 2D (sur une feuille de papier ou un écran), en fonction de leurs similarités perçues. Les résultats sont récoltés sous forme de coordonnées cartésiennes et servent à une Analyse Factorielle Multiple (*Multiple Factor Analysis*), pour extraire encore une fois les attributs propres à justifier cette cartographie. En résumé, l'avantage de cette dernière catégorie de méthodes réside dans la possibilité de mettre à jour des facteurs qui ne pourraient être que difficilement révélés par des évaluations verbales, permettant de positionner les stimuli dans un espace perceptif qui a du sens pour les sujets. Néanmoins, la difficulté d'interpréter ces espaces et les axes qui les constituent constitue la limite principale de ces méthodes.

Des listes d'attributs sonores

[Berg et Rumsey, 1999] discutent de la possibilité de sélectionner les attributs nécessaires et suffisants pour couvrir la totalité des caractéristiques d'un son ou d'un système de reproduction. Il semblerait que le problème soit particulièrement difficile. Certains attributs seraient pertinents dans un cas, mais pas dans un autre. Nous reprenons à notre compte la citation que les auteurs font de [Plomp, 1976] à ce sujet, qui s'exprime sur une expérience ayant utilisé 9 stimuli : « Dans cet exemple, basé sur un ensemble spécifique de stimuli, trois attributs seuls semblent avoir été suffisants pour décrire les différences de façon satisfaisante. Mais ce chiffre ne doit pas être généralisé...il est possible qu'une autre sélection de 9 stimuli ait nécessité, par exemple,

5 dimensions pour représenter leurs timbres de façon précise. » [9]. Le même problème est aussi évoqué et expérimentalement étudié par [Guastavino et Katz, 2004].

Des tentatives ont cependant été menées pour mettre au point des listes d'attributs exhaustives dans le domaine de la qualité sonore. Elles ne prétendent toutefois pas nécessairement au caractère univoque de leurs attributs (i.e., le fait que chaque attribut ait un sens unique), ni à leur orthogonalité (i.e., le fait qu'aucun attribut ne puisse avoir un sens qui se recoupe avec un autre). [Berg et Rumsey, 1999] évoquent même l'utilité d'avoir plusieurs attributs proches de sens, afin de se débarrasser d'erreurs de notation aléatoires localisées sur un ou deux attributs, et d'accéder à ce qu'ils appellent une « variable latente ». Notons toutefois qu'à l'inverse, [Pedersen et Zacharov, 2015], revenant sur les caractéristiques désirables pour un attribut, y évoquent précisément l'indépendance et l'unicité. Pour aller plus loin sur le caractère équivoque de certains attributs, on peut se référer par exemple à [Berg, 2009], qui montre que le mot « enveloppement » (*envelopment*) change de signification entre différentes expériences. Sur la non-orthogonalité des attributs, on peut également se référer à [Guastavino et Katz, 2004], qui recensent plusieurs travaux traitant de l'interdépendance d'attributs.

Parmi les tentatives pour répertorier les attributs, [Letowski, 1989] propose le système MURAL (MUltilevel auditoRy Assessment Language), fondé sur des travaux antérieurs. Il s'agit d'une représentation des attributs sur un disque (voir Figure 4.2). Les attributs sont divisés en deux catégories principales, les attributs timbraux et les attributs spatiaux, avec une sous catégorie à cheval sur les deux. Plus récemment, [Le Bagousse, Paquier et Colomes, 2014] proposent également une liste d'attributs fondée sur un choix d'experts, puis raffinée par positionnement multidimensionnel d'une part, et par un classement libre suivi d'un partitionnement de données d'autre part. De même que pour Letowski, les catégories timbrale et spatiale sont constituées, en plus d'une troisième relative aux défauts de signal. Enfin, les travaux initiés par [Pedersen, 2005] et poursuivis à travers plusieurs publications [Pedersen et Zacharov, 2008 ; Pedersen et Zacharov, 2015 ; Zacharov, Pedersen et Pike, 2016], s'appuient sur plusieurs centaines d'attributs pour proposer une « roue des sons » (voir Figure 4.3) qui réunit 43 attributs, divisés en 8 catégories : attributs de timbre, attributs spatiaux, liés aux artefacts, de balance timbrale, de dynamique, d'intensité et de transparence (avec une huitième catégorie contenant le seul attribut de clarté, déjà présent chez Letowski, à mi-chemin entre un attribut timbral et spatial). On constate ici aussi l'importance accordée au timbre, à l'espace et aux artefacts, qui englobent à eux seuls une trentaine d'attributs.

Attributs du son spatialisé

Nous nous intéressons maintenant au domaine plus circonscrit du son spatialisé. L'ITU recommande, en plus de son unique attribut de qualité audio de base, deux attributs propres à décrire les « systèmes multivoies » [ITU-R, 2015a ; ITU-R, 2015b] [10] : la

9. *In this example, based upon a specific set of stimuli, three factors alone appeared to be sufficient to describe the differences satisfactorily. This number cannot be generalised...it is also possible to select nine stimuli which would require, for example, five dimensions to represent their timbres accurately.*

10. se référer aux versions françaises pour le vocabulaire

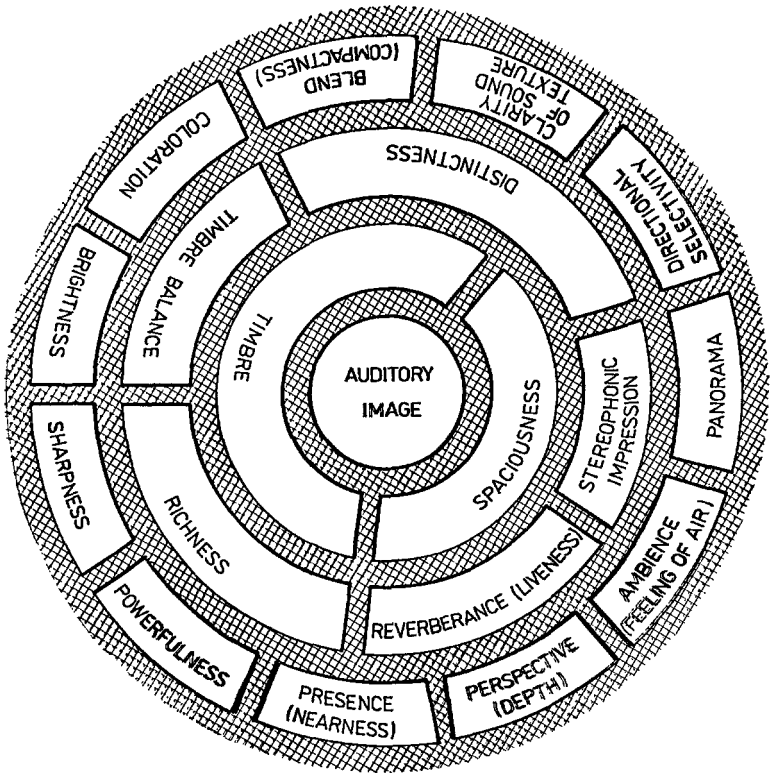


FIGURE 4.2 – Représentation MURAL de [LETOWSKI, 1989]. Les attributs en périphérie sont indépendants mais complémentaires, et sont hiérarchiquement reliés aux catégories sur le même rayon. Plus on s'approche du centre, plus la catégorie est générale.

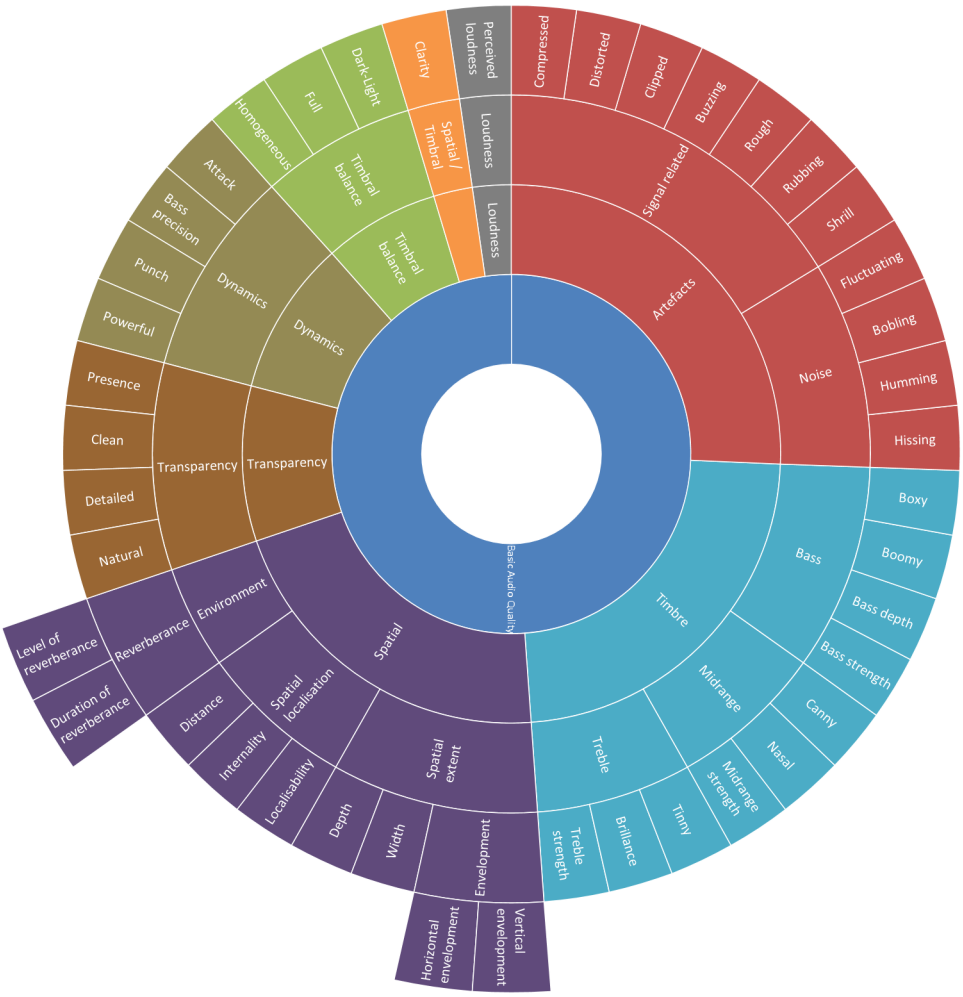


FIGURE 4.3 – Représentation de la roue sonore de [ZACHAROV, PEDERSEN et PIKE, 2016]. De la même façon que pour la roue de la Figure 4.2, les attributs sont en périphérie. Les attributs proches partagent un sens proche. Chaque attribut est hiérarchiquement lié aux catégories qui sont sur le même rayon que lui. Plus on s'approche du centre, plus la catégorie est générale.

qualité frontale de l'image, associée donc à la localisation des sources sonores frontales, et qui porte sur la qualité de l'image stéréophonique (précédemment définie comme relative à l'emplacement des images sonores et aux sensations de profondeur et de réalité de l'événement audio) et les pertes de définition ; et la qualité de la sensation ambiophonique, associés à une sensation spatiale, à l'ambiance ou à des effets ambiophoniques directionnels particuliers. Pour des systèmes sonores évolués (incluant également des systèmes de reproduction sonore spatialisée [ITU-R, 2018]), sont toutefois évoquées :

— la qualité du timbre, décrite par deux ensemble de propriétés, la couleur sonore (e.g., l'éclat, le timbre musical, la coloration, la clarté, la dureté, l'égalisation ou la richesse) et l'homogénéité sonore (e.g., la stabilité, la pureté, le réalisme, la fidélité et les nuances) ;

— la qualité de localisation, qui peut être divisée en qualité de localisation horizontale, qualité de localisation verticale et qualité de localisation distance, et éventuellement qualité de localisation sur l'écran et autour de l'auditeur si les sons sont accompagnés d'images ;

— la qualité de l'environnement, considérée comme une extension de la qualité ambiophonique, qui correspond à l'impression d'espace, à l'enveloppement, à l'ambiance, à la diffusivité ou aux effets spatiaux directionnels d'immersion. Elle-même est divisée en qualités horizontale, verticale et distante.

On constate que même pour les sons spatialisés, les attributs sonores sont divisés ici en qualités timbrales et spatiales, comme on l'a vu dans la section précédente. [Letowski, 1989] a proposé une définition de ces deux types d'attributs : le timbre correspond à cette image auditive par laquelle un auditeur juge du caractère spectral du son. Il permet de distinguer deux sons ayant les mêmes hauteur, intensité, durée et représentation spatiale. L'espace correspond à l'image par laquelle il juge de la distribution des sources sonores et de la taille de l'espace acoustique. Il permet de distinguer deux sons ayant les mêmes hauteur, intensité, durée et timbre, et provenant de deux positions différentes. Fait intéressant, [Rumsey et al., 2005], qui comparent des systèmes de reproduction multicanal, vont même jusqu'à évaluer que leur qualité globale est pour 70% expliquée par leur fidélité timbrale, et pour 30% pour leur fidélité spatiale.

Nous proposons un recensement de travaux qui ont étudié les attributs du son spatialisé, sous la forme d'un tableau (voir les Tables 4.2 et 4.3). La colonne « Publication » renvoie à la ou les publications qui ont participé à l'élaboration des attributs. La colonne « Système de reproduction sonore » énumère les systèmes étudiés. « Méthode d'extraction » résume la méthode employée. Enfin, « Attributs ou catégories d'attributs » répertorie les attributs extraits. Certains travaux utilisent une méthode d'extraction indirecte aboutissant à des catégories sans attributs verbaux précis, ou des catégories élaborées à partir d'un très grand nombre d'attributs. Dans ces cas là, nous signalons simplement les catégories identifiées.

On observe des points communs entre plusieurs publications. En particulier la division générale entre attributs timbraux et attributs spatiaux est récurrente, avec une prépondérance d'attributs spatiaux. Ce consensus relatif est remarquable, dans la mesure où chaque expérience s'attarde sur des systèmes de son spatialisé différents (de qualités probablement différentes), présentés avec des stimuli sonores très variés (extraits musicaux, sons d'ambiance, etc.). Néanmoins, comme le mentionnent [Francombe, Brookes et Mason, 2015], ce n'est encore qu'un consensus relatif. Il y a aussi des

Publication	Système de reproduction sonore	Méthode d'extraction	Attributs ou catégories d'attributs
[T. Nakayama et al., 1971]	multicanal, allant de 1 à 8 canaux	Jugement de préférence et analyse par positionnement multidimensionnel	Attribut timbral : **clarté** (*clearness*) Attributs spatiaux : **plénitude de l'espace** (*fullness*), **profondeur des images source** (*depth of image sources*)
[Berg et Rumsey, 1999]	multicanal, allant de 1 à 5 canaux	*Repertory Grid Technique*	**authenticité/naturel** (*authenticity/naturalness*), **positionnement latéral/taille de la source** (*lateral positioning/source size*), **enveloppement** (*envelopment*) et **profondeur** (*depth*)
[Zacharov et Koivuniemi, 2001c ; Zacharov et Koivuniemi, 2001b ; Zacharov et Koivuniemi, 2001a]	multicanal avec ou sans élévation, allant de 1 à 8 canaux, transaural	*Quantitative Descriptive Analysis*	Attributs timbraux : **richesse** (*richness*), **dureté** (*hardness*), **emphase** (*emphasis*), **couleur de ton** (*tone color*) Attributs spatiaux : **sensation de direction** (*sens of direction*), **sensation de profondeur** (*sense of depth*), **sensation d'espace** (*sense of space*), **sensation de mouvement** (*sense of movement*), **pénétration** (*penetration*), **distance aux événements** (*distance to events*), **largeur** (*broadness*), **naturel** (*naturalness*)
[Berg et Rumsey, 2002]	multicanal 5.0	Attributs pourvus (de [Berg et Rumsey, 1999]) et extraits par *Repertory Grid Technique*	Attributs généraux : **contenu basse fréquence** (*low frequency content*), **naturel** (*naturalness*), **préférence** (*preference*), **présence** (*presence*) Attributs liés à la source : **largeur d'ensemble** (*ensemble width*), **localisation** (*localisation*), **enveloppement de la source** (*source envelopment*), **largeur de la source** (*source width*), **distance de la source** (*source distance*) Attributs liés à la salle : **enveloppement de la salle** (*room envelopment*), **taille de la salle** (*room size*), **niveau de la salle** (*room level*), **largeur de la salle** (*room width*)
[Guastavino et Katz, 2004]	Expérience 1 : multicanal 6.0, 6.1, 12.0 et 12.1 Expérience 2 : multicanal 2.1, 6.1 et 12.1	*Free Choice Profiling*	Expérience 1 : **lisibilité** (*readability*), **présence** (*presence*), **distance** (*distance*), **localisation** (*localization*), **coloration** (*coloration*) et **stabilité de l'image** (*stability of the image*) Expérience 2 : **présence** (*presence*), **lisibilité** (*readability*), **son arrière** (*rear sound*) et **distance** (*distance*)
[Lorho, 2005a]	amélioration spatiale de la stéréo sur casque	*Quantitative Descriptive Analysis*	Attributs de localisation : **sensation de distance** (*sense of distance*), **sensation de direction** (*sense of direction*), **sensation de mouvement** (*sense of movement*), **ratio de localisabilité** (*ratio of localizability*) Attributs spatiaux : **qualité d'écho** (*quality of echo*), **quantité d'écho** (*amount of echo*), **sensation d'espace** (*sense of space*), **équilibre de l'espace** (*balance of space*), **largeur** (*broadness*) Attributs timbraux : **séparabilité** (*separability*), **couleur de ton** (*tone color*), **richesse** (*richness*), **distorsion** (*distortion*), **perturbation** (*disruption*), **clarté** (*clarity*), **équilibre des sons** (*balance of sounds*)
[Lorho, 2005b]	amélioration spatiale de la stéréo sur casque	*Free-Choice Profiling*	Attributs timbraux Attributs spatiaux Attributs liés aux basses fréquences
[Choisel et Wickelmaier, 2006]	multicanal, allant de 1 à 5 canaux	*Repertory Grid Technique* et *Perceptual Structure Analysis*	Attributs spatiaux : **largeur** (*width*), **enveloppement** (*envelopment*), **élévation** (*elevation*), **distance** (*distance*), **spaciosité** (*spaciousness*) Attribut timbral : **éclat** (*brightness*) Attributs reliés ni au timbre ni à l'espace : **naturel** (*naturalness*), **clarté** (*clarity*)
[Geier et al., 2010]	système WFS synthétisé sur casque en binaural	*Repertory Grid Technique*	Attributs de localisation Attributs timbraux

Table 4.2 – Vue d'ensemble de différentes études relatives à l'élaboration d'attributs du son spatialisé.

Publication	Système de reproduction sonore	Méthode d'extraction	Attributs ou catégories d'attributs
[Spors et al., 2013]	divers systèmes multicanaux	Attributs pourvus (synthèse d'attributs de l'état de l'art)	Attributs timbraux : **fidélité timbrale** (*timbral fidelity*), **coloration** (*coloration*), **timbre/couleur de ton** (*timbre/color of ton*), **volume/richesse** (*volume/richness*), **éclat** (*brightness*), **clarté** (*clarity*), **distorsion/artefacts** (*distorsion/artifacts*) Attributs spatiaux : **fidélité spatiale** (*spatial fidelity*), **spaciosité** (*spaciousness*), **largeur** (*width*), **largeur d'ensemble** (*ensemble width*), **enveloppement** (*envelopment*), **profondeur** (*depth*), **distance** (*distance*), **externalisation** (*externalization*), **localisation** (*localization*), **robustesse** (*robustness*), **stabilité** (*stability*)
[Lindau et al., 2014]	Aucun (uniquement fondé sur le vocabulaire)	*Quantitative Descriptive Analysis* et *Repertory Grid Technique*	Attributs timbraux Attributs sur le caractère tonal Attributs sur la géométrie Attributs sur la pièce Attributs sur le comportement temporel Attributs de dynamique Attributs sur les artefacts Attributs généraux
[Francombe, Brookes, Mason et Woodcock, 2016 ; Francombe, Brookes et Mason, 2017 ; Francombe, Brookes, Mason et Woodcock, 2017]	multicanal allant de 1 à 22 canaux, mono basse qualité, mélange de stéréo et binaural sur casque, Ambisonic	mélange de *Free Choice Profiling* et de *Repertory Grid Technique*, puis *Quantitative Descriptive Analysis*, puis *Internal Preference Mapping*	Attributs de large effet : **quantité de distorsion** (*amount of distorsion*), **bande passante** (*bandwidth*), **qualité de sortie** (*output quality*), **enveloppement** (*envelopment*), **largeur horizontale** (*horizontal width*) Attributs d'effet subtil : **profondeur de champ** (*depth of field*), *phasiness*, **basse** (*bass*), **équilibre spatial** (*spatial balance*), **niveau de réverbération** (*level of reverb*), **résonances spectrales** (*spectral resonances*)

Table 4.3 – Vue d'ensemble de différentes études relatives à l'élaboration d'attributs du son spatialisé (suite).

différences : certains attributs apparaissent ou disparaissent selon les publications ; d'autres changent parfois de catégorie. Par exemple, la plupart du temps, clarté et naturel sont respectivement classés comme attribut timbral et attribut spatial, mais rangés dans une catégorie à mi-chemin entre timbral et spatial chez [Choisel et Wickelmaier, 2006] (ce qui au passage était déjà le cas pour la clarté dans la roue des sons Figure 4.3). Autre exemple de changement, celui obtenu par [Guastavino et Katz, 2004], qui proposent deux expériences au protocole identique, mais qui changent les systèmes de restitution et les stimuli, et en extraient des attributs légèrement différents. Un travail d'agrégation de tous ces travaux serait intéressant, permettant de réunir les attributs, d'éliminer les redondances entre les termes qui se recouvrent, et de déterminer le caractère contextuel de certains d'entre eux (par exemple un attribut associé à un système de reproduction précis, ou à un certain type d'auditeurs). C'est le travail que semblent avoir entrepris [Zacharov, Pedersen et Pike, 2016], qui font cependant état d'une recherche encore en cours.

Enfin, en guise de conclusion, on peut remarquer la singularité des catégories de [Berg et Rumsey, 2002]. Les attributs ne sont plus propres à l'ensemble de la scène, mais peuvent être spécifiques aux éléments qui la composent (la source ou la salle). C'est dans la continuité de ces travaux que [Rumsey, 2002] propose sa conception orientée objet de la scène sonore. Les sources peuvent être considérées individuellement, groupées à différentes échelles, jusqu'à la scène jugée dans son entièreté (avec aussi l'environnement indépendamment des source sonores). Les attributs sont alors distribuées aux objets, et sont distingués les micro-attributs (pour les sources individuelles) et macro-attributs (pour un groupe de sources ou la scène entière), permettant d'éviter les confusions quant à leurs sens. Il s'agit sans doute là d'une conception intéressante pour évaluer une scène sonore, car elle permettrait de prendre en compte son caractère évolutif, avec l'arrivée et le départ de certains objets. Elle serait une façon de replacer la scène dans un contexte temporel, première étape vers sa contextualisation, que nous aborderons dans la section 4.4 sur la qualité d'expérience.

Attributs du son binaural

Jusqu'ici nous n'avons pas recensé de travaux qui traiteraient spécifiquement des attributs du son binaural. [Katz et Nicol, 2018] soulignent la difficulté de les lister exhaustivement, en particulier en procédant par une méthode d'analyse dimensionnelle. Cette méthode consisterait à proposer à des sujets de comparer des stimuli entre eux, puis d'analyser statistiquement les comparaisons. Or cela nécessiterait de générer un large ensemble de stimuli capable de représenter une variété de situations auditives. Mais dans le contexte du binaural, les HRTF étant propres à un individu, cette tâche serait particulièrement ardue : un même contenu sonore, filtré par des HRTF individuelles, aboutirait à des stimuli différents pour chaque sujet. Par ailleurs, les HRTF non-individuelles étant perçues de façon différente par chaque sujet, leur utilisation aboutirait à des constructions mentales de l'espace sonore également différentes, rendant l'analyse des résultats particulièrement complexe.

Malgré ces barrières théoriques, [Katz et Nicol, 2018] décrivent un cas d'étude (déjà évoqué dans [Nicol, Emerit, Edwige Roncière et al., 2016]) utilisant précisément une méthode de positionnement multidimensionnel. Le but est de sélectionner des

attributs propres à décrire la qualité de 46 jeux de HRTF non-individuelles (issus de la base LISTEN [Warusfel, 2003]). Le contenu audio utilisé est un extrait en binaural synthétisé à partir d'un enregistrement en 5.1, une émission de Radio France qui consiste en une interview d'un chef de restaurant au milieu d'une cuisine en effervescence. En accord avec ce qui a été présenté dans la sous-section sur les attributs extraits, la méthode consiste en premier lieu à demander aux sujets de juger la dissimilarité entre les stimuli présentés successivement. Plutôt que de présenter les stimuli par paires, la méthode de *Similarity Picking with Permutation of References* est utilisée [Michaud et al., 2013]. Les stimuli sont ici présentés par groupe de 4, 1 référence et 3 stimuli à juger. Le sujet doit sélectionner celui qu'il perçoit comme étant le plus proche de la référence. En présentant successivement toutes les combinaisons de stimuli possibles, chaque stimulus sert à un moment ou à un autre de référence. À l'issue de cette phase, une matrice de dissimilarité est calculée et utilisée pour construire un espace de données, dans lequel les stimuli sont placés en accord avec les jugements des sujets. L'analyse multidimensionnelle révèle qu'un espace à quatre dimensions permet une représentation correcte des stimuli. Quatre experts sont alors recrutés pour interpréter les attributs qui décrivent le mieux ces dimensions. De cette interprétation ressortent les quatre attributs suivants : la quantité de contenu en haute-fréquence, la perception de l'espace (i.e., le volume et l'ampleur de la scène sonore), l'équilibre spatial (ou plus précisément le sentiment qu'une scène sonore est ou n'est pas équilibrée entre la gauche et la droite), et la profondeur ou la distance des sources sonores.

Dans une autre étude, [Simon, Zacharov et Katz, 2016] présentent également une sélection d'attributs adaptée spécifiquement aux HRTF non-individuelles. Pour l'élaborer ils emploient une méthode faisant intervenir des sujets par groupe, et se déroulant selon les trois étapes suivantes : 1) chaque participant sélectionne d'abord individuellement une liste d'attributs après avoir écouté les stimuli par paires ; 2) les listes agrégées sont ensuite discutées par les participants regroupés, avec l'objectif de réduire le nombre d'attributs sur la base d'un consensus ; enfin 3) un test d'évaluation des stimuli à l'aune des attributs est mené, aboutissant à un dernier filtrage. Les contenus audio utilisés sont choisis pour être représentatifs d'une utilisation écologique du binaural (écartant par là les traditionnelles salves de bruits blancs) : un documentaire radio, une musique électronique et une fiction radio. Sept jeux de HRTF sont soigneusement sélectionnés dans la base de données LISTEN [Warusfel, 2003], aboutissant à un total de 21 stimuli, correspondant à 63 paires de stimuli à comparer lors de la première étape de sélection. Cette première étape aboutit à l'évocation de 162 attributs, réduite à la deuxième étape à 12 attributs. Enfin, la dernière étape permet de réduire encore à 8 attributs[11] : modifications spectrales (correspondant à la richesse spectrale du son), élévation, externalisation (perception du son localisé à l'extérieur de la tête), immersion (sentiment d'être localisé à l'intérieur de la scène audio), position avant arrière, position latérale, réalisme (sentiment que les sons proviennent de sources réelles positionnées autour de soi), profondeur du champ sonore (distance entre le son le plus proche et le plus éloigné).

Pour finir, nous pouvons citer de nouveau [Katz et Nicol, 2018], qui donnent à la volée et à l'appui de la littérature quelques pistes sur des attributs propres à qualifier pertinemment le binaural. Ils évoquent par exemple toute propriété physique ou mathématique permettant de décrire la source sonore, sa localisation ou sa localisabilité, sa largeur, l'espace acoustique ou le système de reproduction. De façon intéressante, ils

11. en français dans l'étude

élargissent également à d'autres types d'attributs, qui ne représentent plus seulement des informations purement acoustiques, mais la façon dont l'état psychologique de l'auditeur est affecté par le son. [Nicol, Gros, Colomes, Noisternig et al., 2014], qui évoquent les mêmes attributs, forment justement deux catégories : les attributs physiques d'un côté (*physical-related attributes*), et les attributs psychiques et affectifs de l'autre (*psychic and affective attributes*). Si l'on se réfère aux couches d'abstraction cognitive telles qu'évoquées par [Blauert et Jekosch, 2012] (et ici en Section 4.3.1), tous ces attributs sortiraient donc de la catégorie de la qualité auditive, majoritairement représentée jusqu'ici. [Katz et Nicol, 2018] évoquent le naturel d'une scène, déjà mentionné à plusieurs reprises plus haut, mais dont il est dit qu'il doit être considéré avec prudence, puisqu'il repose sur une référence interne du sujet, inaccessible à l'expérimentateur. D'autres attributs sont proposés, comme la fidélité, la plausibilité d'un stimulus, sa lisibilité, le sentiment de présence, l'immersion ou même l'émotion qu'il suscite (on note d'ailleurs que l'immersion et le réalisme sont deux des attributs trouvés par l'étude de [Simon, Zacharov et Katz, 2016]). Néanmoins, nombre de ces attributs ne seraient plus seulement dépendants du stimuli et du sujet, mais également du contexte qui les environne. Cet aspect de dépendance sera notamment abordé dans la section 4.5, consacrée aux facteurs d'influence de la qualité d'expérience.

4.3.3 Les méthodes d'évaluation

Il existe deux façons possibles d'évaluer la qualité sonore : globalement ou par attribut [Blauert et Jekosch, 2012]. Dans les deux cas, que l'objet du jugement soit le stimulus dans son entièreté ou un seul attribut, les méthodes d'évaluation employées semblent être les mêmes. Vraisemblablement, établir une passerelle entre une note globale et une série de notes d'attributs n'est pas simple. [Blauert et Jekosch, 2012] nous disent qu'il faut être prudent avec l'hypothèse qu'une somme pondérée des notes par attribut permettrait d'obtenir l'équivalent d'une note globale, et qu'« un jugement syncrétique pourrait embrasser plus que la somme des composantes discrètes ». Il vaut mieux donc considérer ces deux méthodes comme servant des objectifs distincts.

[Letowski, 1989] dit à propos des méthodes globales qu'elles peuvent servir à évaluer trois critères : la fidélité d'un son, en comparaison d'un autre, le naturel d'un son, en comparaison d'une référence interne, ou le caractère plaisant d'un son, en comparaison de plusieurs références internes. Mais ces méthodes pourraient tout aussi bien être appliquées directement aux attributs, pour en évaluer les trois mêmes critères. C'est d'ailleurs ce qu'il se passe dans de très nombreuses publications.

Ces méthodes sont largement documentées par l'Union Internationale des Télécommunications (abrégé en ITU en anglais). Celle-ci en recommande plusieurs, selon le type de stimulus et le contexte d'évaluation, et recommande en général de recourir à des échelles de notation. Dans [ITU-R, 2015a], on propose une méthode pour évaluer les petites détériorations de qualité dues à l'encodage d'un échantillon sonore, en le comparant à d'autres via la méthode *double-blind triple-stimulus with hidden reference* (triple stimuli en double aveugle avec une référence cachée) : un échantillon A qui est l'échantillon de référence explicite, et deux échantillons B et C, présentés dans un ordre aléatoire, dont un est la référence cachée et l'autre est l'échantillon dont on souhaite évaluer la qualité. Le sujet, dont on recommande qu'il soit expert, doit évaluer

les dégradations perçues de B et C par rapport à A, via une échelle de notation à 5 degrés (échelle de Likert), allant de « 1 - très ennuyeux » à « 5 - imperceptible ».

Dans la recommandation [ITU-R, 2015b], dédiée à l'évaluation de systèmes audio de qualité intermédiaire, la méthode MUSHRA est introduite, ou méthode dite *double-blind multi-stimulus with hidden reference and hidden anchors* (méthode multi-stimuli en double aveugle avec une référence cachée et des ancres cachées). On recommande ici toujours des sujets avec une bonne expertise d'écoute. À côté d'une référence explicite, des échantillons sont à noter sur des échelles de notations entre 0 et 100, avec 5 paliers allant de « 1 - mauvais » à « 100 - excellent ». Parmi les échantillons se trouvent une référence cachée et deux ancres, l'une basse (i.e., échantillon fortement dégradé) et l'autre moyenne (i.e., échantillon moyennement dégradé).

De façon alternative, [Le Bagousse, Paquier et Colomes, 2012] proposent un test MUSHRA adapté à l'évaluation d'attributs. L'expérience est menée ici sur un système de reproduction 5.1. L'absence de référence explicite permet d'évaluer la qualité plutôt que la fidélité, et les ancres sont dégradées spécifiquement pour chaque attribut. Dans leur expérience, les attributs évalués sont le timbre, l'espace et la présence d'artefacts. L'ancre du timbre est un filtre passe-bas à 3.5 kHz ; l'ancre de l'espace correspond à l'inversion du canal avant droit avec le canal arrière gauche ; et l'ancre des artefacts correspond à l'ajout de bruit rose sur chacun des canaux. À chaque évaluation d'attribut, les 3 ancres sont présentées. Le même test est proposé pour du contenu binaural dans [Le Bagousse, Paquier, Colomes et Moulin, 2011], où l'ancre spatiale correspond à un inversement des canaux gauche et droite par portion, et quelques passages en mono.

Dans [ITU-T, 1996], des méthodes de notation plus simples sont proposées à l'attention de sujets naïfs pour évaluer la qualité de transmission d'un son sur téléphone. La méthode *Absolute Category Rating* (ACR) tout d'abord, qui propose au sujet une note d'opinion moyenne (MOS) sur la qualité perçue d'un son immédiatement après l'avoir écouté, sur une échelle discrète à 5 échelons, allant de « 1 - mauvais » à « 5 - excellent ». La méthode *Degradation Category Rating* (DCR) propose de comparer deux stimuli A et B, A étant toujours la référence explicite, et B l'échantillon à évaluer en comparaison. Une échelle de notation discrète à 5 échelons est proposée, allant de « 1 - la dégradation est très ennuyeuse » à « 5 - la dégradation est imperceptible ». Enfin la méthode *Comparison Category Rating* (CCR) est identique à ceci près que A n'est plus systématiquement la référence : les deux échantillons sont présentés dans un ordre aléatoire. Il faut toujours noter B par rapport à A, mais l'échelle de notation change alors et passe de 5 à 7 échelons, allant de « -3 - bien pire » à « 3 - bien meilleure ».

Chaque type d'échelle est présentée comme étant adaptée à des paramètres expérimentaux différents (sujets experts ou naïfs, différences plus ou moins perceptibles entre les stimuli, etc.) [Zielinski, Brooks et Rumsey, 2007] mettent cependant en garde sur l'utilisation d'échelles avec des échelons purement verbaux. En effet, la distance qui sépare les mots n'est pas toujours uniforme d'un échelon à l'autre, et peut même varier d'une langue à l'autre. Une expérience menée par [Zielinski, Brooks et Rumsey, 2007] démontre cet état de fait, dont la Figure 4.4 illustre les résultats.

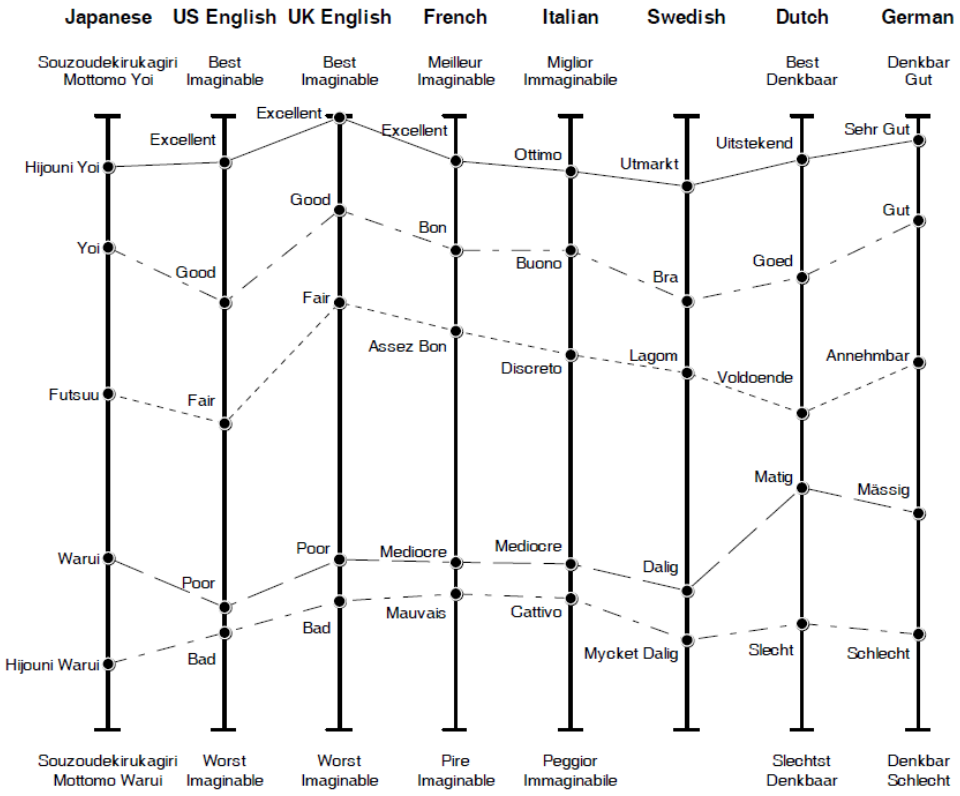


FIGURE 4.4 – Illustration des distances qui séparent les échelons verbaux de l'échelle de notation recommandée par l'ITU, dans différentes langues. Figure tirée de [ZIELINSKI, BROOKS et RUMSEY, 2007].

Pour un récapitulatif sur les échelles de notation, il convient de se reporter à la recommandation [ITU-R, 2019], qui en résume les différents types (discret ou continu, avec ou sans échelons, échelons numériques, verbaux ou les deux, échelles bipolaires, etc.) et leurs emplois dans les autres recommandations.

4.3.4 Le binaural améliore-il la qualité sonore ?

Une fois les attributs et les méthodes d'évaluation passés en revue, nous nous intéressons maintenant aux résultats. À notre connaissance, peu de travaux sur la qualité sonore ont été consacrés au binaural. Plus encore, il est difficile d'en trouver qui comparent le binaural à d'autres systèmes de reproduction, dans le but d'évaluer son apport en termes de qualité ou d'attributs de la qualité. Nous présentons cependant quelques travaux qui s'approchent peu ou prou de cette problématique.

Modalité auditive seule

Dans [Le Bagousse, Paquier, Colomes et Moulin, 2011], une expérience consiste à évaluer des extraits en binaural non-individualisé traités par différents codecs. Les sujets doivent d'abord fournir une note de qualité globale, puis une note de qualité à trois attributs, le timbre, l'espace et le caractère défectueux du son (attribut « défaut »). La procédure consiste en un MUSHRA légèrement modifié (décrit dans la section 4.3.3). 7 extraits sonores sont utilisés, 3 en binaural natif et 4 issus d'une synthèse binaurale d'un son 5.1. Chaque extrait est décliné en 8 versions : le fichier original, le fichier dégradé timbralement, dégradé spatialement, dégradé par ajout de bruit blanc (ces 3 dernières constituent les ancres), et le fichier encodé via 4 codecs différents. Les résultats révèlent une qualité sonore plus élevée pour les fichiers originaux que pour les autres versions. Fait intéressant, les notes attribuées à l'ancre spatiale sont nettement moins bonnes pour les stimuli en binaural natif que pour les stimuli en binaural synthétisés depuis le 5.1. C'est un signe positif pour le binaural natif, puisque l'ancre étant conçue pour susciter un jugement négatif, elle joue alors pleinement son rôle, indiquant une impression de spatialité mieux perçue par les sujets avec ces stimuli. La même expérience est menée sur un système 5.1, mais l'ancrage spatial n'étant pas fait de la même façon, et surtout les codecs testés étant différents, il est difficile de comparer les résultats.

[Nicol, Gros, Colomes, Roncière et al., 2016] comparent différents systèmes d'enregistrements sonores (couples stéréophoniques, six arbres multicanaux (4.0, 5.0, 8.0, etc.), microphones Ambisonics d'ordre 1 et 4, microphones d'appoint spatialisés artificiellement et trois têtes artificielles) tous rendus en binaural non-individualisé. Ici ce sont donc les méthodes d'enregistrement plutôt que les systèmes de restitution qui sont évaluées. Dans un premier test, les sujets doivent reproduire sur une feuille de papier millimétré les sources sonores. Dans le second test, les sujets doivent faire part de leur préférence entre les différents systèmes. Les stimuli sont présentés par paires, tour à tour chaque système présenté avec le stimuli correspondant à la tête artificielle KU100, qui sert de référence. Le sujet donne sa préférence sur une échelle à 7 degrés. Associé à ce jugement, 10 attributs lui sont soumis (crédibilité/réalisme, immersion

sonore, ampleur de la scène sonore, équilibre spatial, externalisation, précision spatiale, coloration, transparence/respect du timbre, effet de salle/réverbération, relief), pour chacun desquels il doit dire si oui ou non il a guidé son choix. Pour les résultats du premier test, les dessins des sujets sont comparés avec un dessin de référence. Les micros d'appoint spatialisés par synthèse binaurale rencontrent le consensus, ainsi qu'un mélange arbre multicanal + micros d'appoint et deux têtes artificielles sur trois, dont la localisation est la plus précise. Pour le second test, la tête artificielle KU100 est majoritairement et significativement préférée, alors qu'elle était dans le bas du classement du premier test. La différence de résultats entre les deux tests confirme que l'efficacité à localiser d'un système n'est pas nécessairement représentatif de sa qualité. En ce qui concerne les attributs pour la préférence, l'immersion sonore, la précision spatiale et le respect du timbre sont les plus mis en avant.

Modalité audiovisuelle

Quelques travaux s'intéressent aussi à l'apport du binaural dans un contexte audiovisuel. Dans [Larsson, Vastfjall et Kleiner, 2002], deux expériences sont menées, mêlant son binaural et casque de réalité virtuelle. Le son binaural est synthétisé en temps-réel et s'adapte aux mouvements de la tête du sujet. Celui-ci doit accomplir une tâche de localisation (trouver des balles dans un environnement d'église) et de mémorisation (retenir des phrases prononcées à chaque fois qu'une balle est trouvée). À l'issue de chaque expérience, le sujet doit remplir un questionnaire relatif à son sentiment d'immersion et à sa mémorisation. Dans la première expérience, sont comparées une scène purement visuelle et la même scène avec du son binaural. Les résultats montrent un sentiment de présence et une mémorisation significativement meilleurs pour la scène avec binaural. La performance de localisation (mesurée par la durée de recherche des balles) n'a pas montré de différence significative entre les deux conditions. Dans la seconde expérience, sont comparées une scène audiovisuelle avec du son stéréo et la même scène avec du son binaural. Les résultats montrent un meilleur sentiment d'immersion avec le binaural, mais pas de meilleure mémorisation, ni de meilleure performance de localisation. Il est intéressant de noter que [Larsson, Vastfjall et Kleiner, 2002] mesurent ici la mémorisation du sujet. Si l'on se resitue dans la perspective de la définition de la qualité sonore selon [Blauert et Jekosch, 2012], citée en section 4.3.1, les attributs énoncés jusqu'à maintenant étaient toujours rangés dans la première ou deuxième couche d'abstraction. La mémorisation, puisqu'elle demande une attention portée sur la sémantique du stimulus, constituerait un attribut de la qualité sonore appartenant à la dernière couche d'abstraction.

Dans [Gonot, Emerit et Château, 2006], des sujets doivent naviguer dans une ville virtuelle à la recherche de neuf sources sonores à collecter. Quatre configurations sont testées, combinant alternativement un son rendu en stéréo ou un son rendu en binaural non individualisé, et une présentation décontextualisée ou contextualisée de l'information (i.e., la distance séparant le sujet de la source est présentée respectivement en coordonnées polaires, ou en « longueur de chemin », tenant compte des routes praticables dans la ville, à la manière d'un GPS). Plusieurs informations sont enregistrées, comme le temps passé par le sujet à un croisement, la distance parcourue ou les changements d'orientation. À la fin de la tâche, le sujet remplit un questionnaire où il évalue une série d'attributs sur une échelle bipolaire graduée (allant d'« absolument » à « pas du tout » en 7 intervalles), relatifs à la qualité sonore, la facilité à utiliser le

son pour naviguer, l'engagement, l'amusement, l'immersion, la cohérence du son avec l'environnement, la facilité à localiser les sons, la facilité de la tâche, l'appréciation générale, l'effet sonore 3D et l'appréciation de cet effet. Les résultats indiquent que pour de nombreux critères, aussi bien enregistrés pendant la tâche que récoltés par questionnaire, la condition de binaural avec contextualisation de l'information est la plus favorable. En particulier, le temps de décision des sujets est plus court en binaural qu'en stéréo.

[Moulin, Nicol et Gros, 2012] évaluent et comparent la qualité ressentie par un sujet face à divers stimuli audiovisuels dont le visuel est rendu en 3D stéréoscopique et dont le son est retransmis en 5.1 via différents systèmes de restitution : un système de 6 enceintes, une barre sonore et un casque ouvert avec reproduction en binaural non-individuel. Les stimuli sont différents extraits d'un documentaire, choisis pour couvrir une variétés de situations sonores (sons environnementaux d'intérieur, d'extérieur, dialogues, musique de fond, etc.) Trois sessions permettent aux sujets de se concentrer successivement sur l'évaluation de la qualité vidéo (degré de profondeur visuelle, confort de visionnage), de la qualité sonore (degré de spatialisation sonore, confort d'écoute) et de la qualité audiovisuelle (degré de cohérence entre son et image et degré d'immersion dans la scène audiovisuelle). L'évaluation se fait sur une échelle à 5 degrés dont seules les extrémités sont labellisées. Les résultats montrent que le degré de spatialisation ressenti est significativement favorisé par la restitution sur casque, devant le système d'enceintes 5.1, tandis que la barre sonore est systématiquement en dernière position. Le confort d'écoute est similaire pour tous les systèmes. L'immersion (rangée ici dans la catégorie des attributs audiovisuels) ne paraît pas non plus impactée par le système de restitution, tandis que le degré de cohérence entre son et image ne l'est que faiblement. Par ailleurs, les attributs vidéos ne présentent pas de changement significatif selon le système de restitution.

Dans [Grani et al., 2014], un système CAVE (une salle sur les murs de laquelle sont projetées des vidéos, de façon à donner l'impression d'être plongé dans un environnement de réalité virtuelle) est accompagné alternativement de trois types de stimuli audio : rendus en stéréo, en binaural, et en binaural spatialement incohérent avec le visuel (décalage constant en position et rotation entre les deux). Les sujets doivent faire un aller-retour dans la salle, chaque traversée étant accompagnée d'un rendu sonore différent. À l'issu de chaque session, le sujet doit donner sur une échelle à 7 degrés son appréciation et la cohérence ressentie. Les résultats montrent une appréciation significativement plus grande pour le binaural par rapport à la stéréo et au binaural incohérent.

Dans [Cobos et al., 2015], les recommandations de l'ITU sont scrupuleusement respectées pour évaluer la qualité sonore de systèmes de reproduction (stéréo, 5.1, 7.1, 10.1 et binaural). Les stimuli sont ici encore une fois audiovisuels, le visuel étant présenté sur une télévision HD. Les attributs évalués sont ceux recommandés par l'ITU, et dont nous avons déjà parlé en section 4.3.2, ainsi que des attributs relatifs au son et à l'image provenant d'autres recommandations : la corrélation entre la position perçue du visuel et celle du son et la corrélation entre l'impression spatiale provoquée par l'image et celle du son. Quatre extraits sont notés par les sujets (film, match de foot, film d'animation, clip musical) en utilisant une échelle ACR dans une première session, puis DCR dans une seconde session, où les stimuli sont comparés deux par deux. Les résultats des deux tests montrent que le binaural est un des systèmes de

reproduction les moins appréciés sur l'ensemble des attributs et des contenus (à peu près au même niveau que la stéréo), à l'exception du film d'animation qui suscite une bonne appréciation, probablement grâce à la profusion d'effets sonores localisés. Les auteurs concluent en argumentant que cette mauvaise notation du binaural est probablement due à l'inconfort relatif engendré par le port du casque obligatoire, à la sensation d'un son parfois localisé à l'intérieur de la tête, ainsi qu'au manque de basses par rapport aux autres systèmes avec caisson de basse. Il semble en tout cas que le type de contenu influence fortement la qualité perçue.

En résumé, ces études envisagent différents aspects possibles de l'apport du binaural (performance de localisation – en termes de temps ou de précision –, préférence, immersion, mémorisation), pour des résultats variés selon les stimuli et les systèmes avec lesquels on le compare (purement visuelle, avec du son mono, stéréo, etc.) Toutefois, les études qui établissent l'apport du binaural concernent souvent des applications interactives, impliquant un rapport direct du sujet à l'espace virtuel. Par ailleurs, certains attributs présentent des résultats divergents : un niveau de préférence marquée mais une performance de localisation moins bonne, une meilleure immersion mais une tâche de mémorisation ou de performance non significative. Tout cela nous montre l'intérêt de considérer une évaluation par attributs plutôt que globalement.

4.3.5 Conclusion sur la qualité sonore

Nous avons exposé le concept de qualité, puis celui de qualité sonore. Nous avons vu celle-ci déclinée en divers attributs au fil de l'état de l'art, permettant de se focaliser sur tel ou tel aspect du son. Il serait néanmoins difficile de définir l'ensemble des attributs nécessaires et suffisants pour décrire la qualité sonore. Les attributs doivent être choisis judicieusement en fonction du sujet d'intérêt choisi par les expérimentateurs. En ce qui concerne le binaural, peu de travaux se sont intéressés à son apport à la qualité sonore, notamment par rapport à d'autres systèmes de restitution sur casque (stéréo et mono). Et pour ceux qui ont abordé la question, les résultats sont variés. La raison tient peut-être au fait que les expériences ont placé à chaque fois les sujets dans des contextes audiovisuels différents, avec écran TV, salle ou casque de réalité virtuelle, et des contenus différents. Nous avons considéré que ces expériences étaient pertinentes dans la section sur la qualité sonore, malgré la présence de visuel, dans la mesure où les seules variables étaient précisément les systèmes de reproduction sonore. Toutefois, gardons à l'esprit que la qualité sonore évaluée en présence d'un flux vidéo n'est potentiellement pas la même que pour un stimulus audio seul (voir par exemple [Beerends et De Caluwe, 1999 ; Hollier et al., 1999 ; Rummukainen et al., 2018]). Cela met en lumière le fait que pour évaluer l'apport du binaural dans une application mobile, il est difficile de se contenter de considérer l'aspect sonore uniquement. Pour cette raison, nous présentons dans la section suivante la notion plus vaste de qualité d'expérience, qui a pour ambition de considérer l'ensemble des éléments intervenant lors d'une expérience.

4.4 La qualité d'expérience

4.4.1 Définitions

[Raake et Egger, 2014] donnent la définition de la qualité d'expérience suivante, en s'appuyant notamment sur les travaux menés par [Le Callet et al., 2013] : « degré de délectation ou d'agacement d'une personne au cours d'une expérience impliquant une application, un service ou un système. Il résulte d'une évaluation de la satisfaction de ses attentes et de ses besoins, formulée au regard d'une utilité ou d'un plaisir présumé, conçu à la lumière du contexte, de sa personnalité et de son état actuel » [12]. Cette définition apporte deux nouveaux éléments par rapport aux définitions sur la qualité de la section 4.2 : la qualité est ici formée *au cours de l'expérience*, et surtout elle dépend du contexte, de la personnalité et de l'état actuel de la personne. Non seulement donc la qualité est rattachée temporellement à l'expérience (impliquant qu'une évaluation trop différée n'aboutirait plus à la qualité d'expérience mais à autre chose), mais elle n'est aussi valable que pour cette unique expérience (impliquant que l'évaluation de la même expérience réitérée pourrait aboutir à une qualité d'expérience différente).

[Weiss et al., 2014] approfondissent cette réflexion en distinguant trois types de qualités d'expérience :

1. une qualité d'expérience instantanée, évaluée pendant l'expérience même. Ce terme inclurait aussi l'évaluation de stimuli de très courtes durées, permettant au sujet de se concentrer sur des variations microscopiques des stimuli ;
2. une qualité rétrospective, évaluée à l'issue de l'expérience ;
3. une qualité cumulative, évaluée à l'issue d'une série d'expériences.

La qualité d'expérience instantanée permet de mesurer le sentiment d'un utilisateur à divers moments de l'expérience, et de recueillir à chaud des impressions face à des événements inattendus pour lui, éventuellement prévus par l'expérimentateur (un changement de débit dans le flux audio ou vidéo, un retournement scénaristique d'une séquence, etc.). La qualité rétrospective s'appuie sur la mémoire de l'utilisateur et permet d'apprécier l'expérience dans son ensemble, bien qu'elle soit assujettie à l'influence des effets de primauté et de récence (faisant que les informations reçues en premier et en dernier au cours de l'expérience marquent davantage la mémoire). Enfin, la qualité cumulative permet de synthétiser la qualité ressentie sur plusieurs expériences. Elle est aussi marquée par les effets de récence et de primauté, nécessite plus de temps pour être évaluée, et est potentiellement soumise à des facteurs d'influence entre les sessions, mais a l'avantage d'être rattachée davantage aux caractéristiques réelles de l'entité plutôt qu'à des fluctuations aléatoires liées à une instance unique ou au contexte environnant. Chacun des trois types de qualité d'expérience permet donc de capturer des informations différentes mais complémentaires, avec des biais et des spécificités de mesure qu'il convient de prendre en compte.

Il est à noter que l'ITU propose aussi une définition de la qualité d'expérience. Alors qu'une ancienne définition est encore en vigueur dans certaines recommandations (par

12. *the degree of delight or annoyance of a person whose experiencing involves an application, service, or system. It results from the person's evaluation of the fulfillment of his or her expectations and needs with respect to the utility and/or enjoyment in the light of the person's context, personality and current state.*

exemple dans [ITU-T, 2008]), une nouvelle version s'appuyant aussi sur [Le Callet et al., 2013] a récemment été proposé dans [ITU-T, 2017] : « degré de délectation ou d'agacement d'un utilisateur d'application ou de service » [13]. Il est par ailleurs ajouté que, compte tenu du caractère mouvant de la recherche actuelle sur le sujet, cette définition serait probablement amenée à évoluer. Si on la compare à celle de [Raake et Egger, 2014], elle correspond presque mot pour mot à sa première partie, à ceci prêt que la « personne » devient ici un utilisateur, et que l'entité évaluée peut être une application ou un service, mais plus un système. Pour compléter cette définition, et venir corroborer davantage celle de [Raake et Egger, 2014], une définition sur les facteurs d'influence de la qualité d'expérience est proposée, qui « incluent le type et les caractéristiques de l'application ou du service, du contexte d'utilisation, des attentes de l'utilisateur vis-à-vis de l'application ou du service et de leur satisfaction, de l'arrière plan culturel de l'utilisateur, de ses problèmes socio-économiques, de son profil psychologique, de son état émotionnel, et d'autres facteurs dont le nombre grandira probablement au fil des recherches futures » [14]. On le voit ici, les aspects de contexte, de personnalité et d'état actuel de la personne évoqués dans la définition de [Raake et Egger, 2014] sont explicités à un niveau plus poussé.

Toutes ces définitions de la qualité d'expérience, telles qu'elles ont été présentées ici, tendent à se rapprocher toujours plus de la notion d'expérience utilisateur [Wechsung et De Moor, 2014]. Bien qu'il ne relève pas de cette thèse de brosser un portrait historique des deux termes, ou d'en dresser toutes les différences, rappelons simplement que la qualité d'expérience s'est construite à l'origine sur la qualité de service [Martín Varela, Skorin-Kapov et Ebrahimi, 2014], née dans le domaine des télécommunications, et axée primitivement sur la qualité en tant que « caractéristique d'un système » (voir Section 4.2.1). De façon analogue, la notion d'expérience utilisateur est plutôt le produit du domaine de l'interface homme-machine, s'attachant à mesurer l'appréciation d'un utilisateur face à un système interactif, et reposant sur le terme plus ancien et plus prosaïque d'utilisabilité. L'une pour les télécommunications, l'autre pour l'interface homme-machine, chacune axée sur des systèmes différents, tendent à se rapprocher de plus en plus. La raison tient sans doute à l'accrétion progressive des technologies jusqu'alors séparées (d'un côté les outils de communication, les téléphones, la radio, etc. et de l'autre les outils interactifs, avec en tête de ligne les ordinateurs, dépourvus d'internet à l'époque, ou les consoles de jeu) en machines de plus en plus polyvalentes et de plus en plus connectées. À tel point qu'aujourd'hui, les deux disciplines vont parfois jusqu'à se fondre en une seule et même entité [Hammer, Egger-Lampl et S. Möller, 2018]. Bien que dans la suite de cet état de l'art, nos recherches se soient davantage focalisées sur la qualité d'expérience, nous ne nous interdisons donc pas de les étendre aux autres domaines sus-cités.

4.4.2 Les attributs de la qualité d'expérience

De la même façon que la qualité sonore, la qualité d'expérience peut être considérée comme un agglomérat d'attributs [S. Möller, Wältermann et Garcia, 2014]. Les

13. *The degree of delight or annoyance of the user of an application or service.*

14. *Include the type and characteristics of the application or service, context of use, the user's expectations with respect to the application or service and their fulfilment, the user's cultural background, socio-economic issues, psychological profiles, emotional state of the user, and other factors whose number will likely expand with further research.*

méthodes d'extractions ou de sélection sont similaires, ainsi que les questionnements sur la possibilité de constituer un ensemble d'attributs nécessaires et suffisants pour décrire la qualité d'expérience de façon complète. Il est toutefois intéressant de noter qu'il n'existe que peu d'auteurs qui s'intéressent aux attributs de la qualité d'expérience en tant que telle. [S. Möller, Wältermann et Garcia, 2014] en font partie et s'appuient sur [Le Callet et al., 2013] et différents travaux dans le domaine des télécommunications pour proposer une catégorisation des attributs par niveau (voir Figure 4.5) :

— le niveau de perception directe. Il concerne tous les attributs propres à qualifier sur le moment les flux d'information qui nous parviennent par voies sensorielles. Nombre des attributs de la qualité sonore rentreraient donc dans cette catégorie. Il peut concerner aussi la relation entre différentes modalités, comme la synchronisation entre une image et un son par exemple.

— le niveau de l'action. Il inclut les attributs reliés à la perception de l'utilisateur de ses propres actions. Cela peut inclure des attributs relatifs à l'espace, pourvu qu'ils soient liés à la perception de l'utilisateur de ses propres mouvements. Dans le cas d'un service de communication par exemple, cela inclut les attributs liés à la parole de l'utilisateur, comme l'écho, ou l'effet local (*sidetone*), i.e., le fait de s'entendre soi-même avec un décalage temporel substantiel.

— le niveau de l'interaction. Il est relatif aux allers-retours entre le système et l'utilisateur, ou entre l'utilisateur et un autre utilisateur (pour des services de communication par exemple). Cela inclut des attributs comme la réactivité du système, le caractère naturel de l'interaction, ou l'efficacité à communiquer ou à converser.

— le niveau d'une instance particulière du service. Il comprend également la situation sociale et physique d'utilisation. Il concerne donc une instance d'utilisation précise, et inclut des attributs tels que l'intuitivité du système, sa facilité d'apprentissage, et l'efficacité à atteindre un but précis lors de cette utilisation. Ce niveau inclut également des attributs non-fonctionnels, comme la « personnalité » du partenaire d'interaction (humain ou machine), ou son esthétique. À ce niveau, on distingue les attributs hédoniques (liés aux sentiments de l'utilisateur) des attributs pragmatiques (liés aux propriétés et à l'ergonomie du service) (voir par exemple à ce sujet [Hassenzahl et al., 2000]).

— le niveau du service. Il est relié aux utilisations successives d'un même service. Il comprend donc des attributs plus globaux, tels que l'utilité du service dans son entièreté (plutôt que d'une fonctionnalité précise) ou sa praticité.

La diversité des attributs présentés (attributs sur la perception, sur le service, sur l'utilisateur, sur l'interaction, etc.), tout autant qu'une certaine uniformité (plusieurs attributs paraissent proches de sens) montrent la difficulté de proposer un cadre théorique propre à englober n'importe quelle expérience dont on voudrait évaluer la qualité. C'est sans doute la raison pour laquelle il y a peu de travaux qui s'intéressent aux attributs de la qualité d'expérience en général (alors qu'à l'inverse, il y a pléthore d'études sur les attributs de la qualité sonore). Certains attributs sont pertinents dans un cas, d'autres non, et [S. Möller, Wältermann et Garcia, 2014] ajoutent même que certains peuvent n'apparaître que sous certaines conditions temporelles (comme la disponibilité ou l'interruption d'une fonctionnalité par exemple).

Pour étayer leur modèle, [S. Möller, Wältermann et Garcia, 2014] donnent deux exemples de domaine d'application et les attributs qui vont avec. Le premier concerne

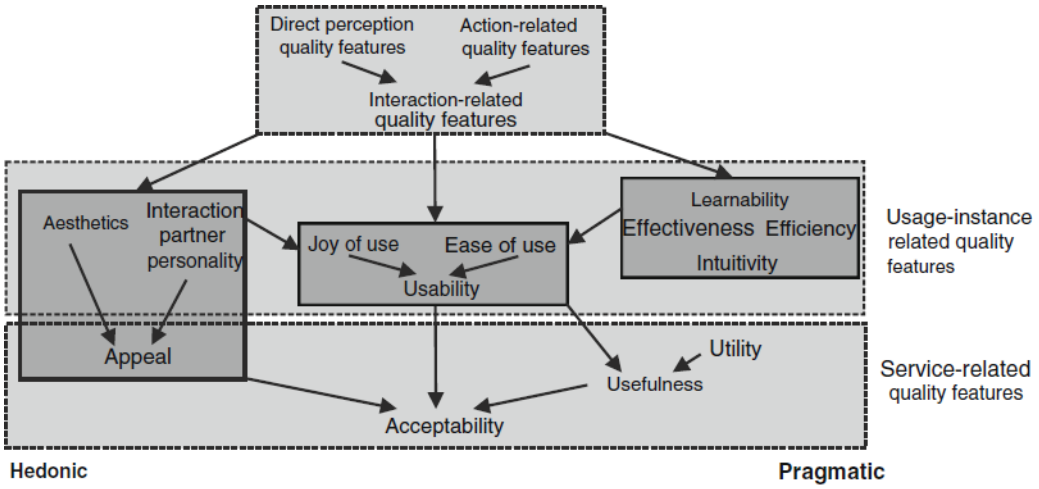


FIGURE 4.5 – Les différents niveaux d'attributs de la qualité d'expérience. Illustration tirée de [S. MÖLLER, WÄLTERMANN et GARCIA, 2014].

les services vocaux, qui ont fait l'objet d'une étude par [WÄLTERMANN, 2013]. À la suite d'une sélection d'attributs fondés sur la littérature, de leur notation par des sujets et d'une analyse statistique, les attributs suivants sont identifiés : la discontinuité (due par exemple à une perte de paquet lors de la transmission de données), le caractère bruité et la coloration du timbre de la voix retransmise. Dans le second exemple, en s'appuyant sur trois études différentes, les attributs de qualité d'expérience pour des services vidéos sont énoncés, à savoir le naturel des couleurs et la netteté perçue [TEUNISSEN et WESTERINK, 1996], la fragmentation, décrivant un certain type de détérioration de l'image [TUCKER, 2011], le sentiment esthétique et le sentiment d'activité [YAMAGISHI et HAYASHI, 2005]. [S. MÖLLER, WÄLTERMANN et GARCIA, 2014] précisent enfin que tous ces attributs appartiennent au niveau de perception directe, et que des méthodes d'identification des attributs des autres niveaux restaient encore à mettre au point.

Peu de travaux s'attachent à recenser les attributs de la qualité d'une expérience sur mobile. La plupart se consacrent davantage aux facteurs d'influence de la qualité d'expérience (voir Section 4.5.1). Les quelques études que nous avons trouvées sont plutôt ciblées sur un type d'application et se concentrent sur les attributs du domaine associé, comme par exemple le visionnage de contenu vidéo [STROHMEIER, JUMISKO-PYYKKÖ et KUNZE, 2010 ; JUMISKO-PYYKKÖ, STROHMEIER et al., 2010 ; CHEON et al., 2015]. Dans cette optique, nous pouvons aussi citer d'autres domaines applicables au support mobile, comme par exemple le jeu vidéo, dont [BEYER et S. MÖLLER, 2014b] listent les catégories principales d'attributs : qualité d'interaction (liée au plaisir de jouer, à la jouabilité, c'est-à-dire à l'efficacité et l'ergonomie du jeu face aux sollicitations du joueur), qualité de jeu (liée à la capacité du joueur à apprendre, contrôler et comprendre le jeu), aspects esthétiques, expérience du joueur (impliquant notamment les notions d'immersion et de *flow* – sensation positive provoquée par un équilibre entre les capacités du joueur et la difficulté du jeu [CSIKSZENTMIHALYI, ABUHAMDEH et NAKAMURA, 2014] –), et acceptabilité (selon sa définition générale, la facilité avec laquelle un utilisateur va utiliser un système ; elle peut être représentée par une mesure purement économique).

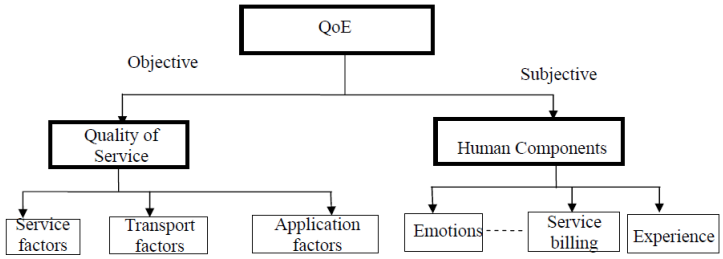


FIGURE 4.6 – Les deux dimensions de la qualité d'expérience selon [ITU-T, 2008]. Les données objectives reposent uniquement sur le service, elles sont indépendantes de l'utilisateur. Les données subjectives concernent le jugement et les émotions de l'utilisateur.

La diversité des applications est telle qu'il serait fastidieux et peu pertinent d'interroger l'état de l'art dans tous les domaines. Gardons simplement à l'esprit que les attributs déjà énoncés dans les Sections 4.3.2 et 4.3.2 sont également ceux considérés du point de vue de la qualité d'expérience lorsqu'on envisage un système sonore spatialisé [FRANK et al., 2014]. Ils garderont donc tout leur intérêt dans un contexte mobile.

4.4.3 Les méthodes d'évaluation de la qualité d'expérience

Les méthodes d'évaluation de la qualité d'expérience sont généralement divisées en deux catégories, qui reposent à l'origine sur la duplicité du terme qualité, telle que nous l'avons exposée en Section 4.2.1. On peut retrouver cette dichotomie par exemple dans l'ancienne définition de la qualité d'expérience donnée par l'ITU [ITU-T, 2008], voir la Figure 4.6. D'un côté, la qualité comprise comme caractéristique d'une entité, à laquelle correspond un ensemble de données indépendantes de l'opinion de l'utilisateur. La première catégorie de méthodes, les méthodes dites objectives, repose sur la collecte et l'interprétation de ces données. De l'autre côté la qualité comprise comme l'expression d'un degré de satisfaction de l'utilisateur. La seconde catégorie de méthodes, les méthodes dites subjectives, repose sur un jugement qu'on aura sollicité de l'utilisateur. Notons que les termes d'objectif et de subjectif sont sujets à débat [15]. [RAAKE et EGGER, 2014] préfèrent parler de méthodes instrumentales et de méthodes basées sur la perception.

Méthodes dites subjectives

Les méthodes subjectives ou fondées sur la perception correspondent aux méthodes déjà présentées dans la Section 4.3.3. Elles font intervenir le jugement de l'utilisateur. Comme pour la qualité sonore, l'évaluation de la qualité d'expérience peut se faire

15. Sur le terme subjectif, [BECH et G. MARTIN, 2005] font par exemple la remarque qu'un sujet à qui on demande de comparer deux cafés, un avec 5 sucres et un sans sucre, percevra certainement que le premier est plus sucré. Les auteurs assignent alors ce jugement à la catégorie des mesures objectives. Sur le terme objectif, le débat est encore plus intense, car l'interprétation des données repose généralement sur des modèles de perception, et sont donc sujets à variation selon les conditions expérimentales.

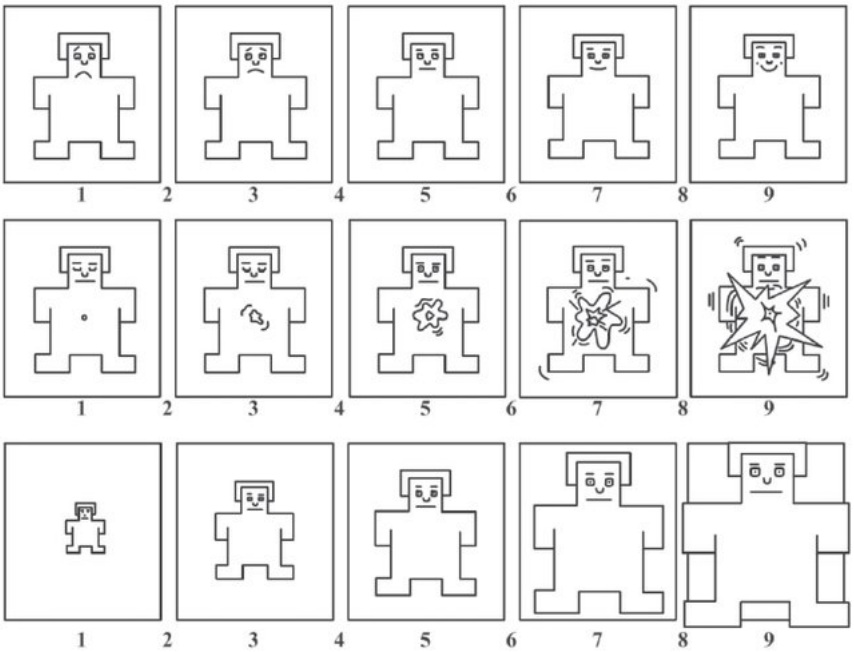


FIGURE 4.7 – Les trois échelles de la méthodes d'évaluation *Self-Assessment Manikin* (SAM) [BRADLEY et LANG, 1994]. La première représente la valence, la deuxième l'excitation et la troisième la dominance.

globalement ou par attribut. On retrouve par exemple les échelles de notation recommandées par l'ITU pour évaluer des stimuli audiovisuels [ITU-T, 1998]. En plus des échelles numériques et verbales, [BRADLEY et LANG, 1994] ont développé un système d'échelles picturales à trois dimensions (voir Figure 4.7), à même de recueillir l'état émotionnel général d'un sujet. Ce système a par exemple été utilisé sur mobile pour mesurer la qualité émotionnelle de rétroactions auditives dans différents contextes [SEEBODE, SCHLEICHER et S. MÖLLER, 2012].

Dans [SCHÖFFLER, 2017], la notion d'expérience globale d'écoute (*Overall Listening Experience*, OLE) est introduite, pour désigner une qualité d'expérience appliquée au domaine audio. L'OLE consiste à demander au sujet : « à quel point avez-vous apprécié écouter cet extrait audio ? », question censée inclure l'ensemble des éléments qui sont rentrés en ligne de compte pour lui, puis de le laisser répondre sur une échelle discrète à 5 degrés, où chaque degré est représenté par une étoile. Le déroulement de l'expérience se divise en deux étapes. Le sujet note d'abord des « items audio basiques », qui représentent les stimuli de référence de chaque extrait utilisé (i.e., les stimuli qui n'ont pas été modifiés par le système de restitution ou le codec à évaluer). De cette façon, ces notes sont censées refléter majoritairement l'appréciation du sujet face au contenu. Tous les stimuli sont présentés sur une même page, à la manière d'un MUSHRA, permettant de revenir sur les notes déjà attribuées, et éviter ainsi des effets de plancher ou de plafond (*floor or ceiling effect*). Ensuite, les conditions à tester sont notées, les stimuli étant désignés comme « items audio ». Une analyse statistique est enfin menée pour comparer les résultats obtenus entre les stimuli de référence et les stimuli à évaluer. Cette méthode récente, employée à plusieurs reprises par son auteur, a aussi été expérimentée ailleurs, par exemple dans [Tim WALTON et EVANS, 2018], que nous détaillons plus bas dans la Section 4.5.2.

Pour noter en particulier la qualité d'expérience instantanée, l'ITU recommande d'utiliser une échelle numérique allant de 0 à 100 [ITU-T, 1998]. Mais lorsque les stimuli engagent la modalité visuelle, le fait de noter en même temps que de visionner peut détourner l'attention du sujet [Pinson et Wolf, 2003]. Ce problème est encore accentué sur un terminal mobile, dont la petite taille empêche de réserver une partie de l'écran à un curseur. [Winkler et Dufaux, 2003] illustrent cette situation, où les auteurs choisissent la couleur du curseur avec soin pour permettre plus facilement aux sujets de l'apercevoir « du coin de l'œil » sur un écran d'ordinateur séparé. [Nagel et al., 2007] proposent pour y remédier une interface de notation sur l'espace des émotions *valence-excitation*, qui permet de superposer l'affichage du stimulus visuel et de l'échelle de notation. Dans la même idée, [Robotham et al., 2018] intègrent dans un environnement de réalité virtuelle une interface d'évaluation MUSHRA pour évaluer des stimuli audiovisuels rendus en binaural. Ils montrent qu'en comparaison d'un MUSHRA classique post-session, seule leur interface permet aux sujets de détecter des dégradations appliquées aux stimuli. D'autres solutions ont été proposées pour se passer purement et simplement d'interface graphique, par exemple [Buchinger et al., 2010], qui présentent un gant pourvu de capteurs de mouvement. Le sujet ferme ou ouvre alors son poing pour aller d'un extrême à l'autre de la notation. [T. Liu et al., 2012] proposent eux d'utiliser un volant de jeu vidéo, que le sujet fait tourner d'un bout à l'autre de sa course pour attribuer une note. Des études complémentaires sont cependant encore nécessaires pour prouver l'efficacité de ces méthodes par rapport au curseur traditionnel et tester la fatigue du sujet sur une utilisation longue durée [Weiss et al., 2014].

Enfin, une autre approche de notation subjective consiste non pas à interroger le sujet sur son ressenti face à un stimulus doté de telle ou telle valeur d'attribut, mais plutôt de le laisser modifier le stimulus en ajustant la valeur à sa convenance. L'analyse consiste ensuite à déceler les différences d'attribution en fonction des sujets, des extraits, des systèmes de restitutions utilisés, etc. On trouve des exemples d'une telle approche dans un nombre grandissant d'études, par exemple dans [Torcoli et al., 2017 ; Shirley et al., 2017 ; Tim Walton, Evans et al., 2018]. Cette méthode convient particulièrement à des scènes audio utilisant le paradigme d'audio orienté objet (*object-based audio*, appelé aussi *scene-based paradigm* dans [Rumsey, 2002], déjà évoqué ici en Section 4.3.2), modèle de représentation de la scène sonore où chaque source est considérée comme un objet, dont les propriétés sont personnalisables indépendamment du reste. Cette représentation est aujourd'hui largement utilisée dans les logiciels de création de jeu vidéo par exemple, où les sources sonores sont placées dans une scène en 3D de façon équivalente à des objets visuels.

Que ce soit pour la qualité d'expérience recueillie au cours de l'expérience ou après, les méthodes subjectives ont la particularité d'exiger de l'utilisateur que le processus de formation de qualité s'opère consciemment. En cela elles sont intrusives et constituent une tâche en soi, qui prend pleinement part à l'expérience. Les méthodes objectives ou instrumentales ne font quant à elle pas intervenir le jugement de l'utilisateur. À ce sujet, [Raake, 2016] confrontent la qualité sonore à la qualité d'expérience en remarquant qu'il est naturel pour la première de faire intervenir une évaluation directe de l'utilisateur, par exemple lorsqu'on compare plusieurs systèmes en vue d'un achat, ou quand on veut explicitement qualifier des systèmes dans un test d'écoute, alors que pour la seconde l'utilisateur est généralement concentré sur le contenu, sans nécessairement en envisager consciemment une évaluation à la fin ([Mausfeld, 2003]

fait référence à ces deux types d'écoute, l'une portée sur le contenu, l'autre sur le contenant – le système de restitution –, comme de la nature duale de la perception). Idéalement, l'évaluation de la qualité d'expérience devrait donc se faire sans interférer avec l'expérience.

Méthodes dites objectives

Les méthodes objectives consistent en la collecte de données relatives à l'utilisation du service, via une tâche explicite ou implicite à accomplir par l'utilisateur, puis à l'interprétation de ces données en termes de qualité d'expérience, en s'appuyant éventuellement sur un modèle de prédiction, permettant de convertir ces données en une valeur de qualité. Les données peuvent être directement issues du service [Rubino, Tirilly et Martın Varela, 2006 ; Wu et al., 2009 ; Menkovski et al., 2009 ; Khan, Sun et Ifeachor, 2012 ; ITU-T, 2015] (ratio signal sur bruit du signal audiovisuel, nombre d'images par seconde, bande passante du réseau, etc.), ou relatives au contexte d'utilisation [Mitra, Zaslavsky et Ahlund, 2015] (température, luminosité, localisation, etc.). L'interaction avec l'utilisateur est aussi source de données utiles à la construction d'un modèle (nombre d'activations des éléments de l'interface, temps de réponse à une sollicitation, temps total d'utilisation, etc.), et particulièrement présente dans le cas d'une application mobile. Si le sujet de l'interaction est maintenant abordé par la communauté de la qualité d'expérience (e.g., [Egger, Reichl et Schoenenberg, 2014]), peu de modèles incluent concrètement ce genre de données (voir néanmoins [Wu et al., 2009] pour un exemple). Enfin notons que les données physiologiques du sujet (rythme cardiaque, sudation, mouvement de l'œil, activité cérébrale, etc.) font également partie des données objectives collectables, et bien qu'encore peu exploitées pour l'élaboration d'un modèle de prédiction de la qualité d'expérience, elles constituent un sujet actif de recherche [Lassalle, 2013 ; Engelke et al., 2016].

Quant au modèle de prédiction de qualité, il s'appuie généralement sur des modèles comportementaux [Wu et al., 2009], des résultats obtenus par méthodes subjectives [Khan, Sun et Ifeachor, 2012], et des algorithmes d'apprentissage automatique [Rubino, Tirilly et Martın Varela, 2006 ; Menkovski et al., 2009].

Dans ces conditions, la collecte de données se fait sans que l'utilisateur n'exprime son opinion quant à la qualité, ce qui place l'évaluation sur une utilisation non-biaisée du service. Cette approche possède également l'avantage de pouvoir être testée dans un contexte hors laboratoire. Néanmoins, son efficacité repose entièrement sur la validité du modèle. Son élaboration nécessitera donc d'une part des validations par des expériences préliminaires, en tenant compte des paramètres de l'application, des potentielles données relatives au contexte, et d'autre part d'une mise à jour à chaque modification de service ou de contexte, opérations coûteuses en temps et en ressources [Brooks et Hestnes, 2010].

Méthodes mixtes

Enfin, notons que quelques expériences combinent les deux types de méthodes. La plupart étudient précisément la corrélation entre les données obtenues par méthode subjective et celles obtenues par méthode objective. [Mauss et al., 2005] par exemple enregistrent des données expérimentales, comportementales et physiologiques mesurant l'émotion de sujets au visionnage d'un film. Le but est de vérifier la cohérence des données entre elles. Premièrement, les sujets doivent reporter continument leur état émotionnel (triste ou amusé) sur une jauge semi-circulaire. Deuxièmement, deux juges mesurent l'expression faciale des sujets, en notant indépendamment l'un de l'autre l'intensité de leurs émotions (amusement – sourire ou rire – ou tristesse – froncement de sourcils, abaissement des lèvres ou pleurs –) moment après moment. Troisièmement, le rythme cardiovasculaire, les activités électrodermale et somatique (i.e., les mouvements du sujet sur sa chaise) sont enregistrés en temps réel. Les résultats indiquent une corrélation significative entre les données des différentes méthodes.

[Jennett et al., 2008] présentent trois expériences visant à mesurer quantitativement l'immersion ressenti par un joueur dans un jeu vidéo. Dans la première, on compare le temps passé par des sujets à compléter un tangram avant et après une session de jeu vidéo. L'hypothèse est que plus les sujets sont immergés dans le jeu vidéo, plus il leur sera difficile d'en « sortir », et plus ils seront lents à résoudre le tangram post-session. La moitié des sujets est confrontée à un jeu considéré comme non-immersif (cliquer sur des cases à cocher qui apparaissent régulièrement à l'écran), et l'autre moitié doit jouer au jeu vidéo *Half-Life* [Valve Corporation, 1998], considéré comme immersif. En toute fin de session, le sujet doit remplir un questionnaire relatif à son sentiment d'immersion. Les résultats montrent une immersion significativement meilleure et un temps de résolution du tangram significativement plus lent pour le jeu immersif. En revanche, l'analyse de corrélation entre les données du questionnaire et celles du tangram est plus incertaine. Une seconde expérience est alors envisagée, où la tâche de tangram est remplacée par une mesure continue du mouvement des yeux (le nombre de fixations des yeux par seconde). Une fois de plus, les résultats montrent des réponses au questionnaire significativement meilleures pour le jeu immersif. Le nombre de fixations par seconde diminue significativement au cours du temps pour le jeu immersif (indiquant la concentration progressive du sujet sur les éléments visuels importants du jeu), tandis qu'il augmente significativement pour le jeu non-immersif (indiquant une perte d'attention progressive). Mais encore une fois, la corrélation entre réponses au questionnaire et mouvement des yeux est plus difficile à établir. Enfin, une dernière expérience évalue par questionnaire l'immersion du joueur uniquement sur le jeu non-immersif, en faisant varier les paramètres d'apparition de la case à cocher. En plus de l'immersion, l'affect émotionnel est recueilli par questionnaire. Les résultats montrent de façon intéressante que les sujets ressentent une immersion relativement haute dans toutes les conditions, bien que certaines déclenchent un affect négatif. Ce résultat met alors en évidence la difficulté de calculer une note de qualité d'expérience globale sur la base d'attributs ciblés, en particulier lorsqu'ils renvoient des informations opposées.

[N. Liu, Y. Liu et X. Wang, 2010] évaluent la qualité d'expérience d'une application de recherche d'emploi en combinant des informations données par l'utilisateur dans un journal de bord (informations contextuelles, fonctionnalités utilisées et retour

utilisateur) et des données d'interaction récupérées via le téléphone (e.g., l'heure d'utilisation ou la liste des icônes sur lesquelles a cliqué l'utilisateur). Si aucune note de qualité d'expérience n'est calculée sur la base des données accumulées, celles-ci servent tout de même de support à une discussion informelle sur les améliorations à apporter à l'application. En particulier, les deux types de données, objectives ou subjectives, fournissent des informations complémentaires : les données d'interaction permettent de reporter certains comportements inattendus ou indésirables, et le journal de bord d'en expliquer en partie les raisons.

[Lassalle, Gros et Coppin, 2011] mesurent quant à eux la qualité d'expérience de contenus audiovisuels à l'aide d'une échelle de notation issue de l'ITU (échelle ACR) et de diverses mesures physiologiques (pouls, température, conductance de la peau, suivi de l'œil). Trois contenus sont testés, un clip musical, un documentaire et un extrait de match sportif, chacun présenté dans une version non dégradée, une version dégradée temporellement (désynchronisations momentanées du son et de l'image) ou une version dégradée par des baisses de débit audio ou vidéo. Si les résultats des tests subjectifs vont dans le sens attendu (meilleure note pour les stimuli non dégradés), aucune concordance n'est trouvée avec les données physiologiques. Étant donnée la difficulté à interpréter de tels signaux, et leur potentielle sensibilité à certains facteurs (effet de nouveauté d'un stimulus, type de contenu, etc.), les auteurs enjoignent à concevoir une expérience de ce type avec le plus grand soin, de façon notamment à ce que les participants soient les plus sensibles possibles à l'effet de la dégradation.

Toutes ces expériences mettent en avant la difficulté à construire un modèle fiable fondé sur des données objectives. Cette difficulté est sans doute en grande partie due aux nombreux facteurs d'influence contextuels. Pour conclure cette section, nous pouvons à nouveau citer l'ITU, qui, dans sa mise à jour récente, évoque également ce point en proposant une définition de l'évaluation de la qualité d'expérience [ITU-T, 2017] : il s'agit du « processus de mesure ou d'estimation de la qualité d'expérience pour un ensemble d'utilisateurs d'une application ou d'un service, au moyen d'une procédure dédiée, et en tenant compte des facteurs d'influence (possiblement contrôlés, mesurés ou simplement collectés et reportés). L'aboutissement de ce processus peut être une valeur scalaire, une représentation multidimensionnelle des résultats, ou des descripteurs verbaux. Toute évaluation de la qualité d'expérience devrait être accompagnée d'une description des facteurs d'influence qui sont inclus. L'évaluation de la qualité d'expérience peut être décrite comme complète lorsqu'elle inclut de nombreux facteurs spécifiques, par exemple une majorité des facteurs connus. En conséquence, une évaluation limitée de la qualité d'expérience ne devrait inclure qu'un seul ou un nombre réduit de facteurs. » [16]. On voit bien ici l'importance soulignée des facteurs d'influence, à l'inverse de l'ancienne définition qui recommande explicitement de ne pas en tenir compte [ITU-T, 2008]. Cette évolution radicale est la preuve d'un changement grandissant à cet endroit depuis plusieurs années, sans doute en partie lié à l'utilisation massive des terminaux mobiles, qui rend les contextes d'expérience multiples et hétérogènes. La section suivante aborde plus en détails les facteurs d'influence de la qualité

16. *The process of measuring or estimating the QoE for a set of users of an application or a service with a dedicated procedure, and considering the influencing factors (possibly controlled, measured, or simply collected and reported). The output of the process may be a scalar value, multi-dimensional representation of the results, and/or verbal descriptors. All assessments of QoE should be accompanied by the description of the influencing factors that are included. The assessment of QoE can be described as comprehensive when it includes many of the specific factors, for example a majority of the known factors. Therefore, a limited QoE assessment would include only one or a small number of factors.*

d'expérience, d'abord d'un point de vue général, puis spécifique au support mobile.

4.5 Le contexte de l'expérience

4.5.1 Contexte et facteurs d'influence de la qualité d'expérience

Les facteurs d'influence de la qualité d'expérience regroupent tout élément qui pourrait intervenir dans l'interaction d'un utilisateur avec un service. Ils sont définis de la façon suivante par [Reiter et al., 2014] (qui reprennent la définition de [Le Callet et al., 2013]) : « toute caractéristique d'un utilisateur, d'un système, d'un service, d'une application ou du contexte dont l'état ou la configuration actuels peuvent influencer la qualité d'expérience de l'utilisateur » [17]. Dans une étude antérieure, [A. K. Dey, Abowd et Salber, 2001] proposent la définition suivante du contexte, très similaire : « toutes les informations pouvant être utilisées pour caractériser la situation d'entités (i.e., une personne, un lieu, ou un objet) et qui peuvent être considérées comme pertinentes par rapport à l'interaction entre un utilisateur et une application, utilisateur et application compris. Le contexte est typiquement le lieu, l'identité et l'état des personnes, groupes et objets informatiques et physiques. » [18]. En regroupant les deux définitions, on peut donc dire que l'ensemble des facteurs d'influence constituent le contexte.

De plus, [Reiter et al., 2014] regroupent les facteurs d'influence en trois catégories : facteurs humains, facteurs système et facteurs contextuels, en précisant toutefois que ces catégories ne sont pas nécessairement exclusives, un facteur pouvant être à cheval sur deux catégories en même temps. Le tableau 4.4 propose un récapitulatif de ces catégories. Par ailleurs, il nous faut bien distinguer le terme de contexte en tant que regroupement de tous les facteurs d'influence de celui de facteur contextuel, qui désigne un facteur qui ne serait ni humain, ni système. À l'exception du terme précis de « facteur contextuel » que nous présentons dans la sous-section ci-dessous, nous emploierons plutôt le mot « contexte » dans son sens le plus englobant, en accord avec la définition de [A. K. Dey, Abowd et Salber, 2001].

Les facteurs d'influence système

Toujours s'appuyant sur [Le Callet et al., 2013] (eux-mêmes inspirés de [Jumisko-Pyykkö, 2011]), [Reiter et al., 2014] proposent la définition suivante : « les propriétés ou caractéristiques qui déterminent la qualité produite techniquement d'une application ou d'un service » [19]. Il s'agit donc de tous les facteurs d'influence liés à

17. *Any characteristic of a user, system, service, application, or context whose actual state or setting may have influence on the Quality of Experience for the user*

18. *Any information that can be used to characterize the situation of entities (i.e., whether a person, place, or object) that are considered relevant to the interaction between a user and an application, including the user and the application themselves. Context is typically the location, identity, and state of people, groups, and computational and physical objects*

19. *Properties and characteristics that determine the technically produced quality of an application or service*

Influence Factors	Type	Examples
System	Content-related	Audio bandwidth, dynamic range ; video motion and detail
	Media-related	Encoding, resolution, sampling rate, frame rate ; synchronization
	Network-related	Bandwidth, delay, jitter, loss, error rate, throughput ; transmission protocol
	Device-related	Display resolution, colors, brightness ; audio channel count
Human	Low-level : physical, emotional, mental constitution	Visual / auditory acuity and sensitivity ; gender, age ; lower-order emotions ; mood ; attention level
	High-level : understanding, interpretation, evaluation	Socio-cultural background ; socio-economic position ; values ; goals ; motivation ; affective states ; previous experiences ; prior knowledge ; skills
Context	Physical context	Location and space ; environmental attributes ; motion
	Temporal Context	Time, duration and frequency of use
	Social Context	Inter-personal relations
	Economic Context	Costs, subscription type, brand
	Task context	Nature of experience ; task type, interruptions, parallelism
	Technical/informational Context	Compatibility, interoperability ; additional informational artifacts

TABLE 4.4 – Vue d'ensemble et exemples des facteurs d'influence potentiels de la qualité d'expérience, selon [REITER et al., 2014].

l'entité utilisée, circonscrite ici à une application ou un service. Quatre sous-catégories sont encore distinguées : les facteurs d'influence système liés au contenu ; ceux liés au média ; ceux liés au réseau ; et ceux liés au matériel. Le premier fait référence au type de contenu, ses caractéristiques, à sa fiabilité et à d'autres attributs indirectement liés. Par exemple, s'il s'agit de données auditives, la bande passante audio et la plage dynamique peuvent être inclus, car elles dépendent fortement du type de contenu véhiculé (e.g., contenu vocal ou musical). Pour des données visuelles, la quantité de détails, la quantité de mouvements, la profondeur des couleurs ou le format (2D ou 3D) sont des exemples typiques de facteurs liés au contenu. Les facteurs liés au média concernent la configuration du média lui-même, son encodage, sa résolution, son taux d'échantillonnage, la fréquence d'images par seconde, la synchronisation entre son et image, etc. Ils sont donc interconnectés avec les facteurs liés au contenu, et peuvent changer en cours d'expérience en fonction des variations des facteurs liés au réseau. La cohérence spatiale entre son et image, telle que nous l'avons étudiée dans la première partie de cette thèse, pourrait sans doute être rangée dans cette catégorie. Les facteurs liés au réseau incluent la bande passante du réseau, le délai, la gigue, le débit, les taux d'erreur et de pertes, etc. Ces facteurs peuvent varier en fonction de la localisation de l'utilisateur, du nombre de personnes sur le même réseau par exemple. Enfin les facteurs liés au matériel font référence au terminal utilisé par l'utilisateur aussi bien qu'au matériel intervenant dans la chaîne de communication (notamment le système de reproduction sonore). Ils désignent l'influence exercée par les avantages et les limitations techniques des appareils, comme la taille de l'écran, sa résolution et la profondeur des couleurs qu'il autorise, la mémoire ou la batterie de l'appareil, la performance des serveurs, etc. Il peut y avoir une forte interaction de ces facteurs

avec ceux liés au contenu : si une image en haute résolution est rendu sur un écran basse résolution avec peu de couleurs par exemple, la qualité d'expérience peut s'en retrouver amoindrie.

Les facteurs d'influence humains

[Reiter et al., 2014] donnent la définition des facteurs d'influence humains : « toute propriété ou caractéristique, variante ou invariante, d'un utilisateur humain. La caractéristique peut décrire son arrière-plan démographique ou socio-économique, sa constitution physique ou mentale, ou son état émotionnel. » [20]. Étant reliés à des processus internes à l'utilisateur et à sa subjectivité, il peut être en effet difficile de les qualifier de façon exhaustive et de les quantifier. Dans la pratique, ils ne peuvent donc qu'être pris en compte jusqu'à une certaine limite. Ici encore, les facteurs sont considérés selon deux catégories, intervenant soit à un bas niveau du traitement de l'information chez l'être humain, ou à un haut niveau. Au bas niveau, on considère les dispositions physiques de l'utilisateur (son acuité visuelle ou auditive, son genre, son âge) et des caractéristiques plus variables comme la motivation (au sens d'engouement), l'attention, l'humeur ou les émotions basiques. Les facteurs de plus haut niveau se rapportent à la compréhension des stimuli et aux processus d'interprétation et d'évaluation associés. Ils sont en lien avec les références internes de l'utilisateur, comme par exemple le cadre socio-culturel et éducationnel, la position socio-économique actuelle, ou les croyances. Ces facteurs prennent de l'importance lorsque par exemple la dimension monétaire d'un service est étudiée. Cette catégorie inclut aussi des facteurs plus variables dans le temps, comme les motivations (au sens d'aspiration), les besoins, les préférences, les traits de personnalité, etc. Certains de ces facteurs peuvent notamment être intrinsèques au service, comme par exemple la motivation dans un jeu vidéo.

Il serait intéressant de croiser ces facteurs bas ou haut niveaux avec les attributs de la qualité sonore organisés par couche d'abstraction, selon [Blauert et Jekosch, 2012] (résumé en section 4.3.1).

Les facteurs d'influence contextuels

Ici encore, [Reiter et al., 2014] nous donnent la définition suivante : « les facteurs qui englobent toute propriété sur la situation qui serait propre à décrire l'environnement de l'utilisateur » [21]. En s'appuyant sur [Jumisko-Pyykkö et Vainio, 2010], qui s'intéressent précisément aux contextes d'utilisation d'un terminal mobile, cinq nouvelles sous-catégories de facteurs sont envisagées :

— Les facteurs physiques se rapportent au lieu où se trouve l'utilisateur, son type (intérieur ou extérieur, lieu de travail, de divertissement, foyer, etc.), à l'organisation de l'espace (taille de lieu, présence d'objets ou d'obstacles, etc.),

20. *Any variant or invariant property or characteristic of a human user. The characteristic can describe the demographic and socio-economic back-ground, the physical and mental constitution, or the user's emotional state*

21. *Factors that embrace any situational property to describe the user's environment*

à des considérations environnementales (place calme ou bruyante, luminosité, température), à leur caractère dynamique (notamment dû aux déplacements de l'utilisateur) et à la mobilité de l'utilisateur (assis, debout, en marche, etc.) ;

— Les facteurs temporels incluent l'heure qu'il est, le moment de la journée (matin, après-midi, soir, moment de travail, moment de pause, moment de repas, etc.), le jour de la semaine, le mois de l'année, la saison, l'année, etc., mais aussi la durée d'utilisation du service, de l'application, ou d'une fonctionnalité, la durée d'un contenu, la fréquence d'utilisation ou la synchronisation entre les événements. Les facteurs physiques et temporels sont souvent associés, correspondant à un contexte spatio-temporel.

— Les facteurs sociaux impliquent la relation aux autres personnes pendant l'expérience. Est-ce que l'utilisateur est seul, accompagné de personnes connues, entouré d'inconnus, en quelle quantité, etc. ? Un autre point est la façon dont l'entourage est impliqué dans l'expérience, y compris à distance. Les relations sociales, culturelles ou professionnelles (relation hiérarchique par exemple) sont également des facteurs importants.

— Les facteurs économiques correspondent au fait que l'application soit payante ou non, et au modèle économique le cas échéant (par exemple un jeu vidéo peut être gratuit ou payant à l'achat, proposer des achats plus ou moins nécessaires en cours de partie, proposer un système d'abonnement, etc.) Cette catégorie peut à l'inverse inclure les modalités et le montant des rémunérations de participants à une expérience de recherche.

— Les facteurs de tâche sont déterminés par la nature de l'expérience. Ils peuvent se référer au caractère multi-tâche d'une expérience (e.g., des activités parallèles à l'utilisation du service qui parasitent l'utilisateur), à l'interruption de tâche ou au type de tâche en question.

— Les facteurs techniques font référence à la relation du matériel à l'environnement, comme le fait que l'appareil utilisé soit connecté à d'autres équipements via Bluetooth, NFC ou Wifi. Un autre facteur serait lié à la disponibilité d'autres services que celui qu'on utilise actuellement (e.g., disponibilité de l'application mobile d'un service pendant qu'on utilise la version sur navigateur web, ou d'un réseau pendant qu'on en utilise un autre), ou la disponibilité de services ou d'outils complémentaires (e.g., accès à un papier et un crayon en complément pour la prise de note en complément du service). L'interopérabilité ou l'adaptation d'un service à différents supports sont également à prendre en compte.

4.5.2 Facteurs d'influence et son binaural

Plusieurs études s'attachent à mesurer l'effet des facteurs d'influence sur la qualité d'expérience dans le domaine de l'évaluation du son binaural. Les travaux présentés dans la section 4.3.4, qui comparent l'avantage de la technologie binaurale par rapport à d'autres systèmes d'enregistrement ou de restitution, ainsi que ceux du chapitre 2 sur l'influence de l'individualisation des HRTF sur les performances de localisation, s'intéressent finalement déjà à ce qu'on vient de décrire comme facteurs d'influence systèmes ou humains. D'autres facteurs systèmes qui influencent significativement la perception du binaural sont la présence de réverbération et de *head-tracking* [Begault, Wenzel et Anderson, 2001]. De même, les comparaisons entre sujets experts et sujets naïfs peuvent être considérées comme des études d'un facteur d'influence d'humain. Dans

[Guastavino, 2003] (étude décrite dans [Guastavino et Katz, 2004]), les sujets sont divisés en ingénieurs du son, acousticiens et sujets naïfs, et se voient présenter des enregistrements d'ambiances sonores enregistrées soit avec un microphone Ambisonic, soit une tête artificielle, ou 5 microphones positionnés à des endroits différents. L'étude montre que selon son niveau d'expertise, le sujet n'oriente pas sa préférence selon les mêmes critères : l'ingénieur du son favorise la localisation et la précision des sources, tandis que les autres mettent en avant la distribution spatiale des sources. À noter que l'ITU propose une méthode de sélection des sujets selon leur niveau d'expertise [ITU-R, 2014].

Un autre sujet étudié en rapport avec la restitution sur casque, en particulier dans une utilisation nomade, est celui de l'influence du bruit ambiant sur la qualité. Une solution largement répandue auprès du grand public est celle des casques audio à réduction de bruit active, qui continuent à faire l'objet d'expériences, comme par exemple dans [Ang, Koh et H. P. Lee, 2017] où leurs performances sont montrées comme fortement dépendantes des conditions environnementales d'écoute. Sur ce même thème du bruit ambiant, [Tim Walton, Evans et al., 2018] s'intéressent à l'audio orienté objet (déjà présenté en section 4.3.2 page 66) dans une optique d'utilisation sur mobile. Dans son expérience, le sujet se voit proposer des extraits sonores rendus sur casque, en même temps qu'un bruit ambiant est rendu sur des enceintes extérieures autour de lui (l'expérience se déroule dans une salle de laboratoire). Le sujet peut ajuster deux paramètres sur un écran d'ordinateur : le niveau sonore général de l'extrait et la balance de volume entre les sources sonores considérées à l'arrière-plan de la scène diffusée sur le casque et celles considérées au premier plan. Le but est de savoir si la présence et le type du bruit ambiant influencent ses réglages. Deux bruits environnementaux sont testés, une ambiance de café et une ambiance de station de métro, aboutissant à quatre conditions : pas de bruit, café calme, café fort et métro. Trois extraits sonores sont proposés, un extrait de match de football (foule en arrière-plan, commentaires sportifs au premier plan), un documentaire radio (musique et atmosphère en arrière-plan, narration au premier plan) et un documentaire TV (musique orchestrale et effets d'arrière-plan, narration et effets saillants au premier plan). Les extraits sont rendus sur un casque ouvert, soit en stéréo, soit sur 5 enceintes virtuelles rendues en binaural, combinées avec la réponse impulsionnelle de la salle et un système de *head-tracking*, permettant donc au sujet une exploration active de la scène sonore, avec la possibilité de focaliser son écoute frontale (acuité d'écoute renforcée) sur les sources d'intérêt. Il n'est pas précisé si le binaural est individualisé ou non. Les résultats révèlent une influence significative du type de bruit ambiant, du type de reproduction sur casque et du type d'extrait sur le niveau sonore général et sur la balance arrière-plan/premier plan. En particulier, plus l'ambiance sonore est élevée, plus les sujets augmentent la proportion d'arrière-plan des extraits, révélant une volonté de masquer les bruits ambiants indésirables par les bruits ambiants de l'extrait. L'influence du degré de similarité spectrale des deux est en conséquence discutée. Par ailleurs, des différences inter-individuelles sont constatées, indiquant la nécessité de tenir compte, en plus des paramètres de l'expérience, des préférences des sujets (et sans doute d'autres facteurs). Deux autres expériences sont menées dans la même étude ; elles examinent plus avant les différences inter-individuelles, mais seulement avec des extraits audio rendus en stéréo.

Sur un autre sujet, [Werner et Klein, 2014] évaluent l'influence de la congruence

des propriétés réverbérantes d'une salle avec un son binaural sur le sentiment d'externalisation et la direction perçue des sources sonores. Alors que le test se déroule dans une salle de test donnée, le sujet se voit présenter des stimuli synthétisés en binaural intégrant soit les propriétés de la salle, soit les propriétés d'une autre salle. L'individualisation du binaural est aussi testée (le sujet possède ou non ses propres HRTF), ainsi que la présence visuelle de la salle d'écoute (avec ou sans lumière). Les résultats montrent une influence significative de la congruence entre les propriétés de la salle et du son binaural, indiquant qu'un meilleur sentiment d'externalisation est ressenti lorsque les deux sont corrélés. Il en va de même pour la visibilité de la salle et l'individualisation des HRTF, qui favorisent aussi le sentiment d'externalisation.

Dans une série de deux expériences, [Fiebig, 2015] évalue l'influence du type d'environnement (environnement apparenté au foyer ou salle de laboratoire) sur la perception d'un bruit d'aspirateur. Dans l'environnement apparenté au foyer, le sujet doit regarder la télévision pendant qu'un autre sujet passe l'aspirateur (cette mise en conditions vise à crédibiliser davantage la scène). En laboratoire, le son d'aspiration seul est présenté via un enregistrement binaural sur casque (enregistré dans la pièce de type foyer, dans des conditions similaires à l'autre version – mêmes distances, mêmes opérations effectuées avec l'appareil –. Aucune précision n'est donnée sur le caractère individualisé ou non de la restitution). L'évaluation du son se fait selon une variété d'attributs, relatifs à l'émotion du sujet ou au caractère du son, et tous notés sur une échelle bipolaire à 7 degrés. Les résultats révèlent une influence notable du contexte de présentation sur certains attributs, comme la netteté du son, le caractère terne, et un effet à la limite de la significativité sur l'acceptabilité ou la puissance du son. Les attributs sont mieux notés dans le contexte de foyer. Au-delà des résultats statistiques, une tendance générale est observée dans ce sens. Dans une seconde expérience, on demande à des sujets d'évaluer un son de bouilloire rendu en binaural (là encore sans précision sur son caractère individualisé ou non), sur une échelle ACR à 10 degrés, allant d'« excellent » à « insupportable ». Le son peut être accompagné ou non d'un visuel. La présence du visuel n'influence pas les réponses des sujets de façon significative, mais un effet croisé entre le type de bouilloire et la présence ou non du visuel indiquerait que l'effet pourrait différer selon le type de produit présenté.

Enfin, [Tim Walton et Evans, 2018] ont souhaité évaluer le rôle des facteurs humains sur l'écoute du binaural. Les facteurs humains sont d'ordre démographique (âge, genre, niveau d'éducation), relatifs à leur expérience audio (utilisation du casque, rapport professionnel au son, connaissance du binaural, participation à d'autres expériences d'écoute) et relatifs à leur attitude vis à vis des technologies audio (compétence, enthousiasme, tendance à se diriger vers des technologies innovantes). Dix extraits musicaux sont utilisés, représentatifs d'une large variété de genres, et rendus en stéréo, en synthèse binaurale (dépendant des conditions d'enregistrement des extraits, et sans précision sur le caractère individualisé ou non), en mono (condition de dégradation spatiale), et en stéréo filtrée avec un passe-bas à 3,5kHz (condition de dégradation timbrale). Les sujets procèdent au test à distance via une interface web, dans leur propre environnement et avec leur propre matériel (avec la préconisation de faire le test dans un environnement calme, avec le même matériel du début à la fin). La procédure consiste à donner les informations relatives aux facteurs humains, puis écouter et évaluer les stimuli, en attribuant à chacun une note d'expérience d'écoute globale (*Overall Listening Experience, OLE*, voir Section 4.4.3), mesurée ici avec la question « à quel point appréciez-vous d'écouter l'extrait musical suivant ? ». La réponse se fait

sur une échelle à 5 degrés allant de « pas du tout » à « beaucoup ». Les stimuli sont présentés par groupe, en concordance avec le protocole d'OLE intronisé par [Schöffler, 2017]. Les résultats montrent en premier lieu une préférence significative de la stéréo sur le binaural, le mono et le filtre à 3,5kHz étant en bas du classement. Les auteurs établissent ensuite quatre types d'auditeurs, selon la corrélation des notes à la qualité spatiale des stimuli, à la qualité timbrale, au contenu ou à la qualité générale. Ils montrent que les facteurs humains ont tous une corrélation significative au type d'auditeur, à l'exception de l'âge, du niveau d'éducation et de l'habitude d'utilisation du casque. En particulier, le facteur compétence (compréhension et connaissance des technologies audio) montre le plus grand niveau de corrélation. Cependant, cette dernière étape sur l'étude de l'influence des facteurs humains exclut le binaural de l'analyse.

Ces quelques études confirment l'intuition que les conditions de présentation d'un stimulus en binaural sont sujettes à modifier la façon dont on le perçoit. De nombreux facteurs d'influence existent ; seuls certains ont été étudiés ici. Ces mêmes facteurs auront peut-être une influence différente dans le cadre d'une expérience sur mobile, et bien sûr, d'autres facteurs interviendront aussi fort probablement. Il s'agira d'identifier dans la suite de ce travail ceux dont il nous faudra tenir compte. Dans la section suivante, nous nous intéressons précisément à différentes expériences relatives aux contextes d'utilisation des terminaux mobiles.

4.5.3 Facteurs d'influence et terminaux mobiles

Les facteurs d'influence revêtent une importance particulière dans le cas des terminaux mobiles, ceux-ci étant conçus pour être utilisés dans des contextes divers et variés [Schleicher, Westermann et Reichmuth, 2014]. Plus encore que les facteurs humains ou systèmes, ce sont les facteurs contextuels qui sont concernés par ce dynamisme (sans toutefois écarter les deux premiers et leurs interactions, e.g., la performance réseau ou l'humeur d'une personne peuvent évoluer avec ses déplacements physiques). Rappelons que la nomenclature des facteurs contextuels proposés ci-dessus par [Reiter et al., 2014] repose sur une publication antérieure, [Jumisko-Pyykkö et Vainio, 2010], qui s'axe précisément sur les contextes d'utilisation d'un terminal mobile.

Si la caractéristique propre aux contextes d'utilisation d'un mobile est la diversité des facteurs d'influence possibles, [Schleicher, Westermann et Reichmuth, 2014] montrent toutefois qu'à l'inverse, d'autres facteurs se retrouvent figés par l'exigence de mobilité : taille de l'objet et poids qui doivent rester bas pour permettre le transport (limitant par la même occasion la taille de l'écran et le nombre d'informations affichables) ; temps d'utilisation limité par la batterie ; recherche de robustesse poussant à réduire les points de fragilité de l'objet, comme par exemple les contrôles physiques (boutons ou roues, dont la suppression est aussi commandée par la volonté de réserver le plus d'espace possible à l'écran) ; restriction d'accès au réseau et à la bande passante. De ce fait, la définition d'un contexte mobile inclurait l'assujettissement de certains facteurs d'influence à des valeurs fixes. Dans la continuité de ces restrictions, une question qu'on peut légitimement se poser concerne les facteurs de localisation et de mobilité. Peut-on considérer des situations d'immobilité, comme le

fait d'être assis ou allongé, et des lieux comme sa maison ou son lieu de travail, comme des contextes mobiles valides ? [Schleicher, Westermann et Reichmuth, 2014] répondent par la positive et établissent par là la définition suivante du contexte mobile : un contexte dans lequel un terminal mobile est utilisé. Cette définition nous semble faire sens, dans la mesure où de très nombreuses situations d'utilisation d'un mobile surviennent quand l'utilisateur ne bouge pas. Les exclure serait sans doute trop limitant dans notre cas. Nous adoptons donc la définition proposée par [Schleicher, Westermann et Reichmuth, 2014] et distinguons par la même occasion le « contexte mobile » du « contexte de mobilité », impliquant lui un rapport au déplacement physique de l'utilisateur (déplacement actif, e.g., la marche, et déplacement passif, e.g., dans un transport en commun).

De nombreux travaux ont tenté d'établir la liste des facteurs d'influence propres aux mobiles, certains plus descriptifs (quels usages, quels facteurs d'influences), et d'autres plus liés à des interactions (quel lien entre la qualité d'expérience et les facteurs d'influence) [Schleicher, Westermann et Reichmuth, 2014]. La majorité se concentre en général sur un type d'expérience ou d'application précis, par exemple la navigation web [Roto et al., 2006], l'utilisation professionnelle du mobile [Wigelius et Väätäjä, 2009], ou les applications grand public [Ickin et al., 2012], tandis que d'autres sont plus généralistes [Korhonen, Arrasvuori et Väänänen-Vainio-Mattila, 2010]. Il a été dit dans la Section 4.4.2 à propos des attributs de la qualité d'expérience qu'il serait long et fastidieux de recenser tous les attributs tant la diversité des expériences était importante. Il en va de même pour les facteurs d'influence sur mobile. D'une certaine façon, tous ceux qui ont été répertoriés dans la Section 4.5.1 pourraient sans doute être pertinents pour un certain type d'application ou d'utilisation. Il serait donc fastidieux d'en faire une liste exhaustive, et cela dépasserait sans doute le cadre de cette thèse. Nous proposons de nous focaliser succinctement sur deux types d'applications possibles, pour lesquelles l'ajout du binaural nous paraît prometteur : le jeu vidéo et l'expérience passive de contenu audiovisuel (film, télévision, etc.)

Dans le domaine du jeu, un sujet qui a particulièrement mobilisé l'attention des chercheurs est celui du *cloud gaming*, permettant de jouer à un jeu vidéo « à distance » depuis n'importe quel mobile, indépendamment de ses performances matérielles[22]. Le principe consiste à faire tourner un jeu sur un serveur distant, de façon à ce que le joueur muni d'une connexion puisse recevoir d'une part les informations du jeu (son et image) et envoyer d'autre part les données d'interaction en temps-réel via le réseau. [Hossfeld, Metzger et Jarschel, 2015] proposent de passer en revue les facteurs d'influence possibles de la qualité d'expérience relative au *cloud gaming*. Ils en distinguent quatre niveaux :

— le niveau du joueur, expert ou naïf (transposé dans le langage usité dans le jeu vidéo en *hardcore gamer* et *casual gamer*). En reprenant la nomenclature de 4.5.1, il s'agit d'un facteur d'influence humain ;

— le niveau du système, avec la latence notamment, induit par la qualité du réseau et l'implémentation du jeu ;

— le niveau du contenu, déterminé en première instance par le genre de jeu, mais aussi par des variables plus bas niveau comme le nombre de décisions ou d'actions

22. Les annonces récentes de Stadia par Google et de Project xCloud par Microsoft (et dans une moindre mesure d'Apple Arcade par Apple), des offres de *cloud gaming* compatibles avec les appareils mobiles, constituent sans doute l'aboutissement le plus éclatant de cette tendance.

requises sur une durée donnée, le temps maximum pour réagir avec succès à des actions dans le jeu, la prévisibilité des actions (un jeu très prévisible sera sans doute moins influencé par la latence par exemple), et l'efficacité et la précision des actions du joueur (aussi bien temporelles que spatiales). Ces deux derniers niveaux concernent des facteurs d'influence systèmes ;

— le niveau du contexte, qui concerne le lieu d'utilisation, le prix du jeu (facteurs d'influence contextuels), mais aussi le cadre socio-culturel du joueur (facteur d'influence humain).

Nombre d'études proposent des expériences pour évaluer l'influence de ces facteurs sur la qualité d'expérience du *cloud gaming*, parmi les plus notables desquelles nous pouvons citer [Jarschel et al., 2011 ; S. Wang et S. Dey, 2012 ; Y.-T. Lee et al., 2012 ; S. Möller, Pommer et al., 2013 ; Beyer et S. Möller, 2014a ; Zadtootaghaj, Schmidt et S. Möller, 2018]. Dans un autre registre, [Pappas et al., 2019] proposent un modèle pour expliquer de quelle façon se combinent différents facteurs d'influence (incluant notamment la qualité du jeu) pour agir sur l'intention d'achat d'un jeu mobile (notion sans doute à rapprocher de la définition de la qualité présumée, présentée en Section 4.4.1). Cependant, à notre connaissance, aucune ne s'est penchée sur l'influence du son dans un jeu vidéo mobile ou sur les facteurs d'influence d'une expérience de jeu vidéo utilisant du binaural.

Dans le domaine de la consommation de contenu audiovisuel, des études montrent également l'importance des facteurs d'influence sur la qualité d'expérience. [Jumisko-Pyykkö et Häkkinen, 2008] évaluent à travers deux expériences l'influence de divers facteurs humains sur la qualité d'expérience pour divers contenus télévisuels visionnés sur mobile (dessin animé, clips musicaux, informations, séries et sport). Les résultats montrent un effet significatif de la plupart des facteurs, en particulier de l'âge et du niveau de relation vis à vis des technologies (situation professionnelle, connaissance des terminaux existants, intérêt pour les nouveautés). [Jumisko-Pyykkö, 2008] montrent quant à eux l'influence significative du débit vidéo, du débit audiovisuel et du ratio de l'image sur la qualité d'expérience du même genre de contenus télévisuels retransmis sur mobile. Ils montrent en particulier l'importance que revêt la qualité audio lorsque les sujets sont confrontés à des débits audiovisuels bas, confirmant des résultats trouvés ailleurs [Winkler et Faller, 2006]. [Catellier et al., 2012] montrent à des sujets des extraits audiovisuels en faisant varier les terminaux mobiles (smartphone, tablette, iPods, et une télévision en tant que référence), impliquant des variations sur la taille de l'écran, la résolution et la distance à l'écran. Les sujets évaluent la qualité sous la forme d'une note d'opinion moyenne. Les résultats indiquent un effet significatif de chaque facteur sur la note obtenue. En particulier, les auteurs concluent qu'une connaissance a priori de l'appareil de diffusion permettrait d'optimiser la bande passante utilisée. L'expérience s'est déroulée dans deux lieux différents, un laboratoire et un simili de pièce à vivre, sans pour autant révéler de différence de résultat entre les deux. [De Pessemier et al., 2012] évaluent de même les facteurs d'influence de la qualité de visionnage de vidéos sur mobile retransmises par flux continu (i.e., en *streaming*). L'expérience se déroule en *Living Lab*, contexte semi-réaliste (voir Section 4.5.4 pour plus de détails), où les participants peuvent regarder les vidéos quand ils veulent et où ils veulent. Deux résolutions vidéos (haute et basse) ainsi que deux transferts de protocoles (via RTP, i.e., un protocole de communication en temps réel et en téléchargement progressif) sont testés, ainsi que divers facteurs d'influence contextuels. Les résultats montrent que les sujets regardent les vidéos majoritairement l'après-midi et le soir, chez eux, et que la présence d'autres personnes dans l'entourage du sujet

ne gêne pas le visionnage. La configuration préférée est la haute résolution alliée au téléchargement progressif, avec une baisse variable de la qualité d'expérience en fonction des dégradations engendrées par le protocole de transfert. De la même façon que pour le jeu mobile, d'autres études récentes s'intéressent au *streaming* [Staelens et al., 2014 ; Hong et Zhu, 2017 ; J. Li et al., 2018]. Ici encore, nous n'avons pas trouvé d'étude plus tournée vers la qualité sonore de contenus audiovisuels sur mobile.

Il semble donc que sur mobile, la tendance générale de la recherche ne s'oriente pas particulièrement vers les technologies de restitution sonore. Malgré tout, les publications évoquées confirment la quantité importante des facteurs d'influence, et l'impossibilité de concevoir une expérience où ils seraient tous maîtrisés. Dans la suite de cette thèse, un travail préalable à toute expérience sera donc de relever les facteurs que nous jugerons les plus saillants, puis de discuter de la possibilité d'en tenir compte, voire de les contrôler. Cette discussion sera l'objet d'une section dans le chapitre suivant.

4.5.4 Évaluation de qualité et contexte, ou la « méthode de déploiement »

Validité interne et validité externe des données

Dans cette section, nous discutons de la notion de validité des données et de la méthode de déploiement d'une expérience de mesure de qualité. Comme il l'est mentionné dans [Scriven, 2005], la validité des données correspond au fait que les résultats d'une expérience permettent de répondre à la question posée initialement par la problématique. Cette notion ne doit pas être confondue avec l'exactitude des données, relative au caractère reproductible d'une expérience. Par exemple, trois expériences successives peuvent reproduire trois fois les mêmes résultats, ce qui leur confère un haut degré d'exactitude. Pour autant leur validité n'est pas assurée. Selon [Scriven, 2005], deux types de validité sont à distinguer : la validité interne, qui reflète le contrôle des facteurs d'influence pouvant biaiser les résultats, et la validité externe, qui représente la mesure dans laquelle le test est représentatif d'un cas d'utilisation réel. La Figure 4.8 schématise les conditions expérimentales requises en fonction du niveau de validité.

Lorsqu'on aborde la qualité d'expérience et les contexte mobiles du point de vue expérimental, la méthode de déploiement et la validité externe des résultats associés sont de première importance. En effet, quelle serait la validité externe d'une expérience mobile menée en contexte de laboratoire ? Tout dépendra de ce qu'on souhaite mesurer. Toutefois, les contextes mobiles étant variés et dynamiques, il y a de fortes chances qu'elle soit faible. [Scriven, 2005] indique qu'un équilibre doit être trouvé entre validités interne et externe selon ce qu'on voudra favoriser, mais qu'un protocole impliquant des sujets conscients de participer à une expérience introduit *de facto* une situation invalide. D'une certaine façon, en rentrant dans l'expérience, le sujet sort du monde réel pour se placer mentalement dans une situation qui ne le fera plus agir d'une façon totalement naturelle, de la même manière qu'on rentre dans le cercle magique d'un jeu [Huizinga et Seresia, 1952].

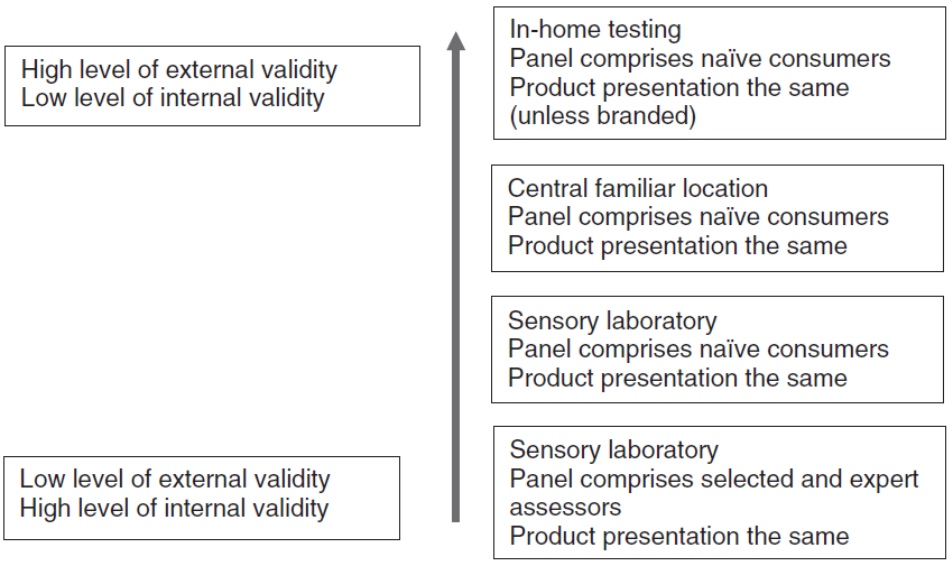


FIGURE 4.8 – Le niveau de validité (interne et externe) en fonction des conditions expérimentales, selon [SCRIVEN, 2005]. Plus la validité externe augmente, plus la validité interne diminue, et inversement.

Le déploiement en milieu contrôlé

Les méthodes de déploiement sont très variées et s'étendent en continu sur le long de l'axe de la Figure 4.8. Au niveau le plus bas de l'axe, il y a les méthodes en laboratoire, qui se déroulent dans une salle de test calibrée, à la façon de ce qui est décrit dans [ITU-T, 2014], où éclairage, humidité, température, réverbération, bruit ambiant, disposition des sièges, etc. doivent être méthodiquement choisis ou réglés. Nombre des expériences décrites dans ce chapitre satisfont en partie ou en totalité à ces critères.

D'autres expériences se déroulent dans des salles de test qui simulent des conditions de vie normale. Nous avons déjà évoqué quelques travaux [CATELLIER et al., 2012 ; FIEBIG, 2015], les seconds montrant d'ailleurs que ce changement de condition influence les données obtenues. Dans le même ordre d'idée, le *Living Lab* est un environnement de test conçu pour simuler une situation de vie semi-réaliste [DE MOOR et al., 2010 ; DE MASI et al., 2016]. Il se définit comme « un environnement pour l'innovation et le développement où des utilisateurs sont exposés à de nouvelles solutions de TIC (Technologies de l'Information et de la Communication) dans des conditions (semi-)réalistes, pour des études moyen ou long terme visant à l'évaluation de nouvelles solutions de TIC et à la découverte d'opportunités innovantes » [FØLSTAD, 2008] [23]. Il y a donc cette notion importante de conditions semi-réalistes, explicitée davantage dans cette définition de [SCHUURMAN, DE MOOR et al., 2011] : « une vraie maison où les activités routinières et les interactions de la vie de tous les jours peuvent être observées, enregistrées pour de futures analyses et manipulées expérimentalement, et où les participants, sujets volontaires de recherche, vivent individuellement, traitant

23. *Environments for innovation and development where users are exposed to new ICT solutions in (semi)realistic contexts, as part of medium- or long-term studies targeting evaluation of new ICT solutions and discovery of innovation opportunities*

le lieu comme leur lieu d'habitation temporaire »[24]. [De Moor et al., 2010] indiquent par ailleurs que le *Living Lab* pour mesurer la qualité d'expérience doit permettre de récolter une variété de données, par l'installation dans les lieux d'outils divers (capteurs, caméras, microphones, etc.), mais qu'il doit aussi respecter quelques contraintes, à commencer par le fait que ces outils ne doivent pas perturber les sujets. Les outils doivent également permettre des mesures à plusieurs niveaux (contexte, social, application, réseau, etc.), doivent être modulaires et permettre l'installation de mises à jour si nécessaire, et doivent fonctionner par protocole IP et être gérables à distance. Si les *Living Lab* offrent des conditions expérimentales rares, alliant le contrôle des conditions expérimentales et la liberté des sujets, ils nécessitent néanmoins d'avoir à disposition des locaux aménagés, disposition particulièrement coûteuse. Pour une vue d'ensemble complète des expériences menées en *Living Labs*, se référer à [Schuurman, De Marez et Ballon, 2015].

Une alternative au *Living Lab*, dont l'intérêt va grandissant, est l'exploitation de la réalité virtuelle, que ce soit dans des systèmes CAVE, déjà évoqués ici dans [Grani et al., 2014], ou via un casque de réalité virtuelle, pour lesquels des études récentes montrent même l'intérêt d'y intégrer directement une interface d'évaluation de type MUSHRA, par rapport à une évaluation plus classique post-session [Robotham et al., 2018].

Le déploiement « hors les murs »

Au-delà des environnements – réels ou virtuels – aménagés pour des expériences, il y a le déploiement en environnement réel. Celui-ci peut être limité, dans la mesure où l'utilisation du service ou du produit est ciblée à un contexte spécifique. Par exemple dans [Shively, 1998 ; K. Beresford et al., 2006], où des tests d'écoute sont menés pour évaluer un système de reproduction sonore en voiture, ou dans [Staelens et al., 2014] où les tests se déroulent d'une façon plus extensive « à la maison », pour mesurer la qualité de contenus audiovisuels délivrés en *streaming* sur tablette. [Schöffler, 2017] développe et propose une interface web de MUSHRA permettant à des sujets de participer à des tests d'écoute depuis n'importe quel ordinateur. L'expérimentateur prend alors bien soin d'adjoindre à son test une fiche d'instruction avec des recommandations sur le contexte idéal, le niveau de bruit ambiant, le type de casque à utiliser, le temps disponible, etc.

Dans des contextes encore plus libres, diverses méthodes existent pour collecter les données. Dans [Jacucci et al., 2007] par exemple, les sujets, successivement spectateurs d'un rallye et d'un festival, sont suivis à la trace par des observateurs qui notent leurs faits et gestes. Bien qu'elle soit un moyen efficace de prendre en compte un grand nombre de facteurs d'influence, cette méthode nécessite un nombre de ressources supplémentaires potentiellement coûteuses. Autres inconvénients : non seulement la simple présence des observateurs peut altérer le comportement des sujets, mais les observateurs eux-mêmes peuvent être sources de biais, ne serait-ce que par la difficulté de bien voir ce qui se passe sur le petit écran mobile des sujets. Pour remédier

24. *A real home where the routine activities and interactions of everyday home life can be observed, recorded for later analysis, and experimentally manipulated, and where volunteer research participants individually live in, treating it as a temporary home*

à ce problème, [Brown, McGregor et Laurier, 2013] remplacent l'observateur humain par une capture continue d'écran, pré-installée sur le téléphone du sujet, doublée d'une caméra portative accrochée sur le sujet. Ainsi, les erreurs de l'observateur et son caractère intrusif sont atténués, mais par pour autant supprimés, la caméra accrochée étant potentiellement très embarrassante sur la durée. [Froehlich et al., 2007] proposent un système permettant de capturer des données d'interaction et de recueillir des informations par questionnaire sur le téléphone. Nous rapprochons cette méthode de celle présentée dans [N. Liu, Y. Liu et X. Wang, 2010], déjà mentionnée dans la Section 4.4.3, mélangeant informations d'interaction directement récoltées depuis le téléphone et journal de bord numérique.

Avec ces dernières méthodes, où le sujet est « lâché dans la nature », la validité externe est aussi conditionnée par la façon dont celui-ci prendra part à l'expérience. Dans le genre d'études qu'on vient de mentionner, les sujets viennent au laboratoire se faire installer une application, ou l'installent à distance via un face-à-face par vidéo interposée avec l'expérimentateur, ou se font même prêter un téléphone qui ne leur appartient pas. Cette étape préalable, qui fait partie intégrante de l'expérience, conditionne le comportement des sujets a priori. D'autres études, mentionnées par [Schleicher, Westermann et Reichmuth, 2014] comme des études à grande échelle, rendent leur application de test téléchargeable directement via une boutique d'application grand public comme le Google Play Store ou l'App Store. Dans ces conditions, on a l'avantage certain de pouvoir recruter un nombre significativement plus grand de sujets. En revanche, la collecte des données est plus limitée, car davantage soumise à des considération de respect de la vie privée, et anonyme. [Gonzalez, Hidalgo et Barabasi, 2008] et [Henze, Rukzio et Boll, 2011] en proposent deux exemples, les premiers en étudiant les trajectoires de 100 000 téléphones sur une durée de 6 mois, les seconds en analysant la précision des interactions tactiles avec l'écran sur plus de 120 millions de données recueillies auprès de plus de 90 000 téléphones. Des outils d'aide au déploiement et à la collecte de données à grande échelle sont proposés par des API telles que celle proposée par [Wagner, Rice et A. R. Beresford, 2013], ou celle utilisée par [Böhmer et al., 2011].

Dans les expériences consistant à observer les usages quotidiens d'un smartphone, les sujets ont pour mission de l'utiliser normalement, sans se poser les questions de quand, de combien de fois, ou de la durée d'utilisation. Mais dans des expériences qui requièrent l'utilisation d'une application dédiée, ou des retours subjectifs réguliers du sujet, ces trois questions sont encore une fois d'importance vis-à-vis de la validité externe. Une expérience proposant des sessions trop longues risquent d'effriter la motivation des sujets ; une expérience avec trop peu de sessions pourrait fournir des résultats limités, etc. Une méthode qui intègre ces aspects au cœur de son protocole est la méthode dite *Experience Sampling Method* (ESM, ou méthode d'expérience par échantillons) [Larson et Csikszentmihalyi, 2014]. La méthode émane du domaine de la psychologie qui a pour sujet d'étude le bien-être dans la vie de tous les jours. Fondée initialement sur le principe du journal intime (catégorie de méthodes dont une présentation générale peut être trouvée dans [Bolger, Davis et Rafaeli, 2003]), la procédure consiste à demander au sujet de reporter lui-même diverses informations relatives à son état, son contexte et ses activités, à des moments aléatoires de la journée et plusieurs fois par jour, afin d'obtenir un ensemble d'échantillons représentatifs de son quotidien. Chaque session ne doit pas prendre plus de quelques minutes, pour minimiser son caractère intrusif. [Larson et Csikszentmihalyi, 2014] proposent une

répartition idéale des sessions : une session aléatoire par bloc de 2h, en commençant la journée à 8h et en allant jusqu'à 22h, pendant une semaine. Mais ils ajoutent qu'il est possible de varier, diminuer le nombre de sessions par jour et augmenter la durée totale de l'expérience par exemple. Une discussion des divers paramètres et de leurs variations peut être consultée dans [Consolvo et Walker, 2003]. Ce protocole sied particulièrement bien aux expériences menées sur mobile, dans la mesure où la totalité du déroulé peut se faire sur le terminal (notifications, utilisation d'applications et réponses aux questionnaires, dont les données peuvent être transmises par le réseau). Cette méthode a par exemple été utilisée par [Traer et J. H. McDermott, 2016], qui récoltent les coordonnées GPS de sujets et leur demandent par SMS des informations subjectives sur l'endroit où ils sont. Ils utilisent ensuite ces informations pour aller mesurer la réponse impulsionnelle des lieux, et analyser le lien entre profil de réverbération de la salle et perception des sujets. Un détail intéressant de leur protocole est que les sujets sont payés à chaque SMS envoyé. Cette façon de faire permet de maintenir la motivation du sujet tout au long de l'expérience.

Dans une autre version, [Ickin et al., 2012] récoltent diverses informations pour analyser les facteurs d'influence de la qualité d'expérience sur des applications grand public. Les informations sont issues à la fois de données du téléphone envoyées automatiquement (force du signal – réseau téléphonique, Wifi ou Bluetooth –, quantité de données transmises et envoyées, nombre d'appels et de messages, luminosité de l'écran, vitesse de réponse d'un serveur via le réseau, etc.), et via un questionnaire sur la qualité d'expérience ressentie et sur le contexte du sujet (lieu, contexte social, niveau de mobilité). Le questionnaire est conçu pour ne pas prendre plus de quelques secondes à répondre. Un point important est qu'il n'est pas proposé systématiquement, afin de ne pas biaiser l'utilisation d'une application par effet d'anticipation. D'autres informations sont collectées en complément une fois par semaine, dans un dialogue direct entre le sujet et l'expérimentateur, pour éventuellement écarter les données aberrantes, établir des relations de causalité entre les différentes données, et de ce fait identifier les facteurs de la qualité d'expérience prépondérants pour chaque sujet.

La méthode ESM convient aux expériences dans lesquelles le moment de l'interaction ou du questionnaire peut être décidé par les expérimentateurs. Mais elle ne convient pas dans le cas où l'interaction survient d'un besoin inopiné du sujet. Dans [Sohn et al., 2008] par exemple, on veut étudier le lien entre le besoin d'une information sur mobile et l'interaction qui en résulte. Étant donné que le besoin d'information ne se commande pas et dépend entièrement du contexte du sujet, la méthode ESM, bien que mentionnée par les auteurs, est écartée au profit d'une méthode où le sujet reporte les informations et données de façon autonome. Cette méthode souffre cependant du fait que les sujets peuvent oublier de reporter. Dans [A. Möller et al., 2013], on compare les données d'interaction d'utilisation d'un smartphone (quelles applications, quelle durée, etc.) avec les données reportées directement par l'utilisateur dans une application dédiée. Le but est de mesurer concrètement la véracité des propos reportés par des utilisateurs dans ces méthodes de type journal de bord. Ils trouvent qu'il y a une différence significative entre les éléments reportés par les sujets et les données d'interaction : les sujets ne reportent en général pas plus de 70% de leur utilisation, et jusqu'à moins de 40%. Ils surestiment aussi significativement la durée d'utilisation d'une application, et oublient fréquemment de reporter l'utilisation s'ils n'ont pas de rappel. Plus important encore, l'auto-report change même l'utilisation qu'ont les sujets de leurs applications. Cette étude montre donc la faiblesse de ces méthodes, de

la nécessité de notifier le sujet quand c'est possible, et de la confiance prudente qu'il faut accorder à leur jugement sur leurs propres utilisations.

Les sujets non consentants et les données déjà existantes

La méthode dont la validité externe serait maximum serait une méthode dans laquelle le sujet n'est pas au courant qu'il participe à une expérience, ou dont les données sont récoltées sans son consentement préalable. Une telle pratique peut évidemment poser des questions juridiques en termes de respect de la vie privée, évoquées non sans ironie par [Schleicher, Westermann et Reichmuth, 2014] à propos des affaires régulières sur la collecte de données gouvernementales ou par de grandes entreprises internationales. Une autre solution simple serait d'analyser des données existantes déjà publiquement accessibles. Par exemple, [Golder et Macy, 2011] utilisent les millions de messages disponibles sur Twitter à travers le monde pour analyser les humeurs à travers les changements quotidiens, hebdomadaires ou saisonniers. Il faut dans ce cas que l'objet d'étude soit compatible avec le genre de données auxquelles on a accès.

Conclusion sur la méthode de déploiement

Plusieurs études mesurent l'influence de la méthode de déploiement sur la qualité d'expérience. Quand différences il y a, elles montrent en général une qualité d'expérience et une tolérance accrue des sujets dans les contextes réalistes. Mais ces études semblent encore isolées et appliquées à des cas d'utilisation de services très précis (voir par exemple [Kaikkonen et al., 2005 ; Barnard et al., 2007 ; Jumisko-Pyykkö et Hannuksela, 2008 ; Liang et Yeh, 2011 ; Jumisko-Pyykkö et Utriainen, 2011 ; Catellier et al., 2012] sur le déploiement en contextes mobiles). Au delà donc de la proposition théorique de [Scriven, 2005] sur la validité externe des données, cette question reste encore largement ouverte. Un double travail à la fois quantitatif et qualitatif sera nécessaire pour lever progressivement le voile sur la pertinence des méthodes de déploiement : d'une part une étude systématique de l'état de l'art, permettant de glaner l'ensemble des méthodes et des résultats obtenus, et dont le recoupement pourra donner des indications, et d'autre part bien sûr d'autres expériences, mettant en concurrence directe plusieurs méthodes pour en comparer les résultats.

4.6 Conclusion sur l'état de l'art

Nous avons présenté un état de l'art sur l'évaluation de qualité. Nous l'avons organisé d'une façon hiérarchique, considérant dans un premier temps l'expérience sonore seule. Nous avons pu présenter la qualité sonore associée, ses attributs, ses méthodes d'évaluation, et les résultats quant à l'apport du binaural. Nous avons ensuite élargi en considérant la qualité d'expérience, qui considère non plus seulement le son, mais la totalité des éléments inhérents à l'interaction d'un utilisation avec un service. Là encore, nous avons considéré les attributs et les méthodes d'évaluation de la qualité d'expérience. Nous nous sommes ensuite penchés sur la question foisonnante du

contexte et des facteurs d'influence de qualité, en nous focalisant plus particulièrement sur l'expérience mobile. Enfin, nous avons abordé les méthodes de déploiement de l'expérience, permettant de rendre compte du contrôle que les expérimentateurs peuvent exercer sur les facteurs d'influence, à la faveur ou au détriment du caractère réaliste des résultats.

Aucun travail ne s'est proposé jusqu'à maintenant d'aborder l'apport du binaural sur mobile. Si la problématique de cette thèse reste donc entière, les sections de ce chapitre nous permettent d'éclairer d'une lumière nouvelle certaines questions posées dans le Chapitre 1. La question des apports possibles du binaural peut maintenant être observée du point de vue des attributs de la qualité sonore, de la qualité d'expérience et de leurs méthodes d'évaluations. Par ailleurs, la question de la méthodologie pour mesurer ces apports dans un contexte mobile est à relier à la méthode de déploiement et aux facteurs d'influence. Le chapitre suivant est l'objet de ce travail. L'objectif sera de dégager les grands axes d'une expérience pour mesurer de l'apport du binaural sur mobile.

Chapitre 5

Vers une expérience sur l'évaluation de l'apport du binaural dans une application mobile audiovisuelle

5.1 Introduction

Ce chapitre a pour objectif de présenter les différentes problématiques de recherche auxquelles nous sommes confrontés dans cette thèse, et d'exposer les solutions que nous envisageons. Dans le chapitre 4, nous avons proposé un état de l'art qui permet d'envisager l'expérience du binaural, d'abord en ne considérant que le son lui-même, et la qualité sonore associée ; puis en replaçant le son au sein d'une expérience mobile audiovisuelle, et en considérant alors la qualité d'expérience ; puis enfin en contextualisant cette expérience et en prenant compte les facteurs d'influence susceptibles d'agir sur la qualité d'expérience.

Aucune expérience ne s'est intéressée jusqu'à maintenant au cas du binaural dans une application mobile audiovisuelle en contexte écologique. Les récents changements de définitions de l'ITU à propos de la qualité d'expérience [ITU-T, 2017] montrent bien qu'on ne peut plus simplement s'appuyer sur leurs anciennes recommandations en termes de méthode, qu'il faut en expérimenter de nouvelles. Aucune recommandation ne propose par exemple de méthode adaptée exclusivement aux tests sur mobile. Du côté du binaural, et plus généralement du son spatialisé, la tendance à considérer le son comme intégré à une expérience globale se fait pourtant de plus en plus présente ; en témoignent par exemple les deux thèse récentes de Michael Schöffler [Schöffler, 2017] et de Timothy Walton [Timothy Walton, 2018]. En particulier, Walton s'intéresse à l'effet du bruit ambiant sur la perception d'une scène sonore rendue en binaural. Il est dit que le but de cette expérience est de se placer dans les conditions d'utilisation d'un terminal mobile. Or l'expérience se passe sur un ordinateur fixe, et le sujet contrôle une interface à l'écran pour évaluer les sons, situation encore loin de représenter l'usage réel. Par ailleurs, les extraits utilisés sont purement sonores et en majorité extraits de contenus audiovisuels dépourvus de leur visuel (extrait audio de match sportif ou extrait de documentaire TV). On sait pourtant l'influence que

peut avoir le visuel sur le ressenti de la qualité audio [Beerends et De Caluwe, 1999 ; Hollier et al., 1999 ; Rummukainen et al., 2018]. Il y a donc une nécessité à envisager des protocoles expérimentaux plus proches de conditions réalistes, pour conférer aux données cette validité écologique mentionnée par [Scriven, 2005] et discutée dans le chapitre précédent. Pour ce faire, notre ambition est de nous inspirer de cette autre partie de l'état de l'art, consacrée aux usages mobiles et aux méthodes de déploiement qui permettent leur observation.

En nous appuyant sur ces trois niveaux d'étude de la qualité – son, expérience, contexte –, le présent chapitre a pour but d'argumenter les axes de recherche vers lesquels nous nous orientons, au travers de trois questions : quel type d'apport du binaural souhaitons-nous évaluer ? Dans quelle expérience mobile, c'est-à-dire sur quelle application ? Et dans quel contexte, c'est-à-dire avec quelle méthode de déploiement ? Tout l'objet de ce chapitre sera de répondre par des choix à ces trois questions et, en ça, de préciser les modalités selon lesquelles nous traiterons via l'expérience du chapitre 6 la quatrième et dernière question, problématique principale de la thèse : quel est l'apport du binaural dans une application mobile audiovisuelle ?

Nous nous proposons de traiter ces questions en discutant d'abord de ce qui conditionne l'expérience dans ses aspects les plus globaux (à l'inverse de l'état de l'art). Nous commençons par la méthode de déploiement et les facteurs d'influence dans la section 5.2, sur laquelle nous nous appuierons pour proposer une application en section 5.3, pour enfin nous pencher sur les attributs du son binaural en section 5.4. Enfin, un bref récapitulatif en section 5.5 nous permettra de peindre une vue d'ensemble de l'expérience présentée dans le chapitre 6.

5.2 Quel rapport au contexte ?

5.2.1 La méthode de déploiement

Le parti pris de cette thèse est de connaître l'apport du binaural tel qu'il serait perçu par le grand public. Selon [Scriven, 2005], nous avantageons donc la validité externe sur la validité interne, i.e., la validité écologique des données, à défaut d'un contrôle total sur les facteurs d'influence. Dans les méthodes qui ont été exposées dans l'état de l'art, les méthodes de type journal intime ont attiré notre attention. S'appuyant directement sur les retours du sujet, en complément d'autres données collectées par le téléphone, elles présentent l'avantage de ne pas nécessiter de mise en place lourde et intrusive, comme des expérimentateurs qui suivraient à la trace leurs sujets, ou du matériel supplémentaire qui leur serait harnaché, comme une caméra ou un microphone. De cette façon, des conditions écologiques plausibles sont conservées, permettant de faire durer l'expérience aussi longtemps que voulu. Par ailleurs, ces méthodes consistent à solliciter le sujet régulièrement, sur de courtes sessions, pendant une longue échelle de temps. Elles permettent ainsi d'accéder à ce que nous avons appelé la qualité d'expérience cumulative (voir Section 4.4.1 du chapitre 4), c'est-à-dire une qualité d'expérience représentative d'une utilisation consistante de l'application, moins soumise aux aléas d'une utilisation particulière. Ici donc, nous sommes certes soumis à des facteurs d'influence divers, sur lesquels nous n'avons pas

de contrôle, et par lesquels la qualité d'expérience sera conditionnée. Cependant, la durée de l'expérience et la juste répartition des sessions doivent nous permettre de rencontrer suffisamment de situations (inter- et intra-sujets) pour dégager une tendance générale.

Parmi les méthodes de type journal intime, la méthode dite *Experience Sampling Method* (ESM) apparaît comme la plus adaptée à nos besoins. Elle ne requiert pas d'attendre que l'utilisateur utilise son téléphone de lui-même. On peut donc décider de quand, combien de fois et pendant combien de temps l'expérience se déroulera. Dans ce protocole, il faudra particulièrement être attentif à l'équilibre entre tous ces éléments, tout en évitant que le sujet se sente submergé. En particulier, pour l'ESM il est recommandé d'accomplir des sessions de courte durée. Cela aura une conséquence non seulement sur l'expérience en elle-même, sachant qu'une exposition à du contenu audiovisuel devra sans doute prendre quelques minutes pour revêtir un caractère réaliste, mais aussi sur la collecte des données subjectives (quantité d'informations demandées, type d'échelle le cas échéant, etc.), le questionnaire ne devant pas prendre trop de temps. Il faudra également prendre soin de notifier le sujet des sessions à accomplir, en vertu des résultats montrés par [A. Möller et al., 2013], où les sujets ne reportent pas plus de 70% de leur activité lorsqu'ils sont livrés à eux-mêmes.

5.2.2 Les facteurs d'influence

En choisissant la méthode ESM, le sujet peut effectuer ses sessions dans n'importe quel contexte. Si nous n'avons aucun contrôle dessus, il est toutefois judicieux de collecter des informations sur les facteurs d'influence pertinents à l'expérience. Toutefois, la diversité et la quantité de ces facteurs est telle qu'il serait impossible de tous les prendre en compte. Il faut donc se concentrer sur ceux pouvant influencer notre perception du binaural d'une part, et qui sont accessibles à la mesure d'autre part. L'inférence du contexte peut se faire de plusieurs façons : par une analyse de données récoltées par les capteurs de l'appareil (gyroscope, microphone, capteur de luminosité, etc.), par une analyse des données d'interaction entre l'utilisateur et le service, ou encore simplement en demandant à l'utilisateur. Les données récupérées automatiquement posent le problème qu'elles requièrent un modèle solide pour être interprétées, modèle susceptible de varier d'un contexte à l'autre. Cela constituerait une étude à part entière. Par ailleurs, récupérer des données de capteurs, tels que le microphone par exemple, pourrait poser de sérieux problèmes d'atteinte à la vie privée, même dans le cadre consenti d'une expérience. Pour cette raison, dans un souci de simplicité et d'efficacité, nous choisissons plutôt d'interroger directement l'utilisateur. Nous orientons alors notre choix sur des questions factuelles qui ne sont pas soumises à l'interprétation du sujet (nous permettant de tirer nous-même les conclusions de l'influence des facteurs), ou des informations perceptives qui seraient difficilement accessibles par un autre moyen.

La question maintenant qui se pose est celle du choix des facteurs d'influence. Le binaural étant une technique de rendu sonore spatialisée, trois facteurs d'influence importants peuvent constituer des sources possibles d'interférence : le niveau de bruit ambiant, la charge mentale du sujet et son rapport à l'espace environnant. Pour le

bruit ambiant, puisque nous faisons le choix de nous appuyer uniquement sur des informations fournies par le sujet, les plus factuelles possibles, nous nous appuyons sur le lieu dans lequel il se trouve (à la maison, au travail/à l'école, dans la rue, en transport, autre intérieur, autre extérieur), et son entourage (seul(e), avec une ou plusieurs personnes connues, entouré(e) d'une ou plusieurs personnes inconnues). L'entourage peut également servir à évaluer sa charge mentale. Pour en apprendre plus à ce sujet, nous lui demandons également son niveau de mobilité (assis(e)/allongé(e), debout, en train de marcher) et une évaluation subjective de sa charge mentale (avec une échelle discrète allant de « 0-complètement libre d'esprit » à « 5-très occupé(e) », avec des intermédiaires purement numériques). Cette dernière valeur subjective nous semble indispensable car difficile d'accès autrement. Des tests préliminaires nous poussent à limiter les questions au nombre de quatre, pour ne pas donner un sentiment de lourdeur au sujet, et maintenir la session dans une durée raisonnable.

Une combinaison astucieuse de ces valeurs peut nous aider à former les contours de contextes mobiles qu'il sera intéressant de comparer : par exemple « seul chez soi, assis ou allongé, avec un faible niveau d'occupation », contexte mobile calme par excellence, et « debout ou en train de marcher, en extérieur, avec un fort niveau d'occupation avec des personnes connues », qui représenterait un contexte hautement perturbant.

D'autres données contextuelles sont collectées au moyen d'un questionnaire de début d'expérience : la marque, le modèle du téléphone et le système d'exploitation, qui peuvent avoir une conséquence sur les performances de l'application, le modèle du casque audio, le genre du sujet, son âge ou ses habitudes d'utilisation de smartphone. Comme ces données ne sont pas propres à chaque session, mais propres à chaque sujet, nous ne pourrons pas les envisager de la même façon que les autres pour une analyse statistique (seulement comme un facteur aléatoire « sujet », global à tous les facteurs). Elles pourront cependant servir à l'observation qualitative des données.

En résumé donc, si on reprend la nomenclature des facteurs d'influence présentée dans la Section 4.5.1 du chapitre précédent, on cumule des facteurs d'influence systèmes (type de téléphone, type de casque, système d'exploitation), humains (âge, genre, habitude d'utilisation du smartphone) et contextuels (deux facteurs physiques avec le lieu et le niveau de mobilité ; un facteur social avec l'entourage ; et un facteur de tâche avec le niveau d'occupation).

De nombreux autres facteurs auraient sans doute été pertinents. Nous l'avons déjà dit : il n'est pas possible de tous les prendre en compte dans une seule expérience. Seule l'accumulation d'études, indépendantes et connexes, dans la continuité de ce que proposent par exemple [Tim Walton, Evans et al., 2018], permettrait de cartographier petit à petit l'influence du contexte sur l'expérience du binaural lorsqu'il est couplé à un terminal mobile.

5.3 Quelle application ?

Le choix de l'application doit être guidé par de nombreuses considérations, parfois théoriques, parfois pratiques, parfois complémentaires et parfois contradictoires. Dans

la section qui suit, nous nous intéressons au domaine de la taxonomie, utile pour catégoriser les applications selon des critères précis.

5.3.1 Une taxonomie des applications

Taxonomie et applications mobiles, définition

La taxonomie est un domaine qui consiste à décrire des objets pour les classifier [Nickerson et al., 2009]. Deux approches existent pour procéder à la classification : la phénétique, lorsqu'on regroupe les objets selon leurs similarités apparentes, et la cladistique, si on les regroupe en regardant leurs processus de construction. En ce qui nous concerne, nous voulons classifier les applications selon certains critères apparents (phénétique), plutôt qu'en fonction de la façon dont elles ont été conçues ou programmées. [Nickerson et al., 2009] séparent les critères de classification en dimensions, chacune faite de caractéristiques « mutuellement exclusives et collectivement exhaustives ». Dans cette publication, les dimensions et caractéristiques suivantes sont proposées :

— Dimension temporelle : l'application propose des interactions en temps-réel ou asynchrone ;

— Sens de communication : application d'information (de l'application vers l'utilisateur), de report (utilisateur vers application), ou d'interaction (dans les deux sens) ;

— Transaction : l'utilisateur peut procéder à des transactions financières via l'application ou non ;

— Dimension publique : application disponible pour tout le monde ou pour certaines personnes seulement (application d'entreprise par exemple) ;

— Multiplicité : l'utilisateur utilise l'application seul ou participe à l'utilisation au sein d'un groupe (jeux multijoueurs par exemple) ;

— Localisation : l'application propose un service qui dépend de la localisation de l'utilisateur ou non ;

— Identité : l'application personnalise son service en fonction de l'identité de l'utilisateur ou non.

Cette taxonomie est conçue pour classifier un grand nombre d'applications selon des modalités de fonctionnement très variées. Plus communément, les classifications utilisées par les magasins d'application (e.g., le *Google Play Store*, l'*App Store* ou le *Windows Phone Store*) permettent de distinguer les applications selon leur thématique principale (Business, Éducation, Jeux, Réseaux Sociaux, etc.), ou d'autres critères comme les plus téléchargées, l'âge recommandé, les applications recommandées sur la base d'anciens achats, etc. Dans le cas de notre recherche cependant, toutes ces classifications sont sans doute trop généralistes (des dimensions qui ne nous concernent pas, comme l'identité ou la transaction chez [Nickerson et al., 2009] par exemple), et à l'inverse pas assez précises, en omettant par exemple le rapport entre son et image, le rapport au contexte, etc. [Nickerson et al., 2009] évoquent le fait que plusieurs travaux procèdent à des taxonomies de façon *ad hoc*, sans nécessairement rechercher l'exhaustivité des dimensions. On en trouve des exemples chez [Kennedy-Eden et Gretzel, 2012] pour les applications de tourisme, chez [Brockmann, Stieglitz et Lattemann, 2014] pour les applications liées au monde de l'entreprise, chez [Gibbs,

Gretzel et Saltzman, 2016] pour les applications dédiées aux hôtels de marque, ou encore chez [Riggs et Gordon, 2017] pour les applications de planification urbaine. Le travail que nous proposons dans la section suivante se situe en quelque sorte dans le sillage de ces taxonomies à usage plus circonscrit.

Proposition de taxonomie

Trois dimensions nous importent pour sélectionner notre application : le rapport de l'application au contexte, le rapport qu'entretiennent le son et l'image au sein de l'application et le mode d'interaction entre l'utilisateur et l'application.

Le rapport au contexte nous intéresse car notre méthode expérimentale a pour particularité de se dérouler dans la vie de tous les jours. Nous distinguons trois types (ou caractéristiques, au sens de [Nickerson et al., 2009]) d'applications au regard de ce critère : les applications *context-aware*, les applications *context-driven* et les applications *context-passive*. Une application *context-aware* adapte son interaction avec l'utilisateur en inférant le contexte qui l'environne. C'est un sujet largement étudié par la recherche [Schilit, Adams et Want, 1994 ; A. K. Dey, Abowd et Salber, 2001 ; Tamminen et al., 2004 ; Adomavicius, 2011 ; Zheng et Jose, 2019]. Les données récoltées servent à adapter les caractéristiques de l'application en fonction de ce contexte déduit, un exemple typique étant d'augmenter la luminosité de l'écran si l'utilisateur est dans un endroit sombre, ou de réduire le bruit ambiant avec un casque à réduction de bruit actif. Un autre exemple peut être consulté dans [Smith, Ma et Ryan, 2006], où l'application interagit avec l'utilisateur par messages textes plutôt que par synthèse vocale si le bruit ambiant est trop important, ou inversement s'il est en voiture. Les applications *context-driven* sont celles dont l'utilisation n'a de sens que pour un contexte précis. Un exemple d'une telle application serait une application de GPS, adaptée uniquement à une utilisation en voiture (pas de sens à l'utiliser hors déplacement, et une précision de géolocalisation, un mode de guidage et une carte pas forcément adaptés à un déplacement à pied ou en transport en commun). Néanmoins, ces applications ne sont pas non plus adaptées à la méthode ESM. Enfin la dernière catégorie concerne toutes les autres applications, dites *context-passive*, qui sont accessibles et fonctionnent de façon équivalente dans tous les contextes.

Les applications *context-driven* ne sont pas adaptées à notre protocole utilisant l'ESM. En ce qui concerne les applications *context-aware*, nous remarquons que le *mapping* entre données de contexte et perception de qualité constitue un problème de grande envergure. Chaque contexte (identifié par les facteurs d'influence que nous récoltons) nécessiterait une étude de qualité permettant de le relier à des paramètres d'application optimaux. Nous percevons la difficulté et la fragilité d'une telle entreprise (reposant dans notre cas sur des informations contextuelles malgré tout limitées), et nous écartons également cette catégorie. En conséquence donc, nous nous intéressons aux applications *context-passive*.

Le rapport entre son et image nous intéresse car il va nous guider dans la façon dont on mettra le binaural en scène. Nous identifions trois types d'association son/image sur terminal mobile. Dès maintenant, notons que ces comportements peuvent se retrouver mélangés au sein d'une seule et même application.

1. D'une part il y a les sons corrélés sémantiquement au rendu visuel : les dialogues, bruitages et musiques lors du visionnage d'un film, les bruits d'interaction avec le menu d'un jeu vidéo, etc. Ces sons peuvent être à leur tour de deux sous-catégories : diégétique et non-diégétique. Diégétique, dit d'un son intégré à l'action, auquel on peut associer une position dans l'espace. Par exemple les voix de personnages qui dialoguent dans un film. Notons que la position spatiale associée au son diégétique n'est pas forcément visible à l'écran. Par exemple, dans un dialogue, la caméra étant fixée sur un seul des protagonistes, on entend à la fois la personne face caméra et la personne hors-champ. Bien que la personne hors-champ soit invisible à l'écran, on lui associe quand même une position dans l'espace, par rapport à la caméra. Ce son est donc diégétique. La musique d'un film est elle un bon exemple de son non-diégétique, car elle ne peut pas être entendue par les protagonistes du film ; elle ne fait donc pas partie de l'action ; elle n'a pas de position spatiale prédéfinie. Elle n'en reste pas moins associée au rendu visuel. Toujours dans cette catégorie de sons associés au visuel, nous pouvons ajouter les sons de navigation dans une interface, comme ceux d'un menu de jeu vidéo par exemple.
2. D'autre part, il y a les sons décorrélés sémantiquement du visuel, une situation particulièrement courante sur mobile : par exemple lorsqu'on écoute une musique, tout en naviguant dans le menu de l'application pour choisir la suivante. La musique n'a pas de lien avec le menu. Ce cas apparaît aussi lors d'une utilisation multi-tâche du mobile, quand par exemple on superpose conversation téléphonique et navigation internet.
3. Enfin, ajoutons la catégorie des applications sans rendu visuel du tout. Caractéristique plus minoritaire, nous la mentionnons néanmoins dans un but d'exhaustivité. Le jeu vidéo *A Blind Legend* [DOWINO, 2015] en est un exemple, dans lequel on dirige un personnage aveugle grâce aux sons qu'il entend. Le jeu utilise le binaural, et si le rendu visuel est absent, il est néanmoins possible d'interagir avec l'écran.

Si chaque catégorie peut contenir des applications susceptibles d'être enrichies par du son binaural, la première retient notre attention, celle du son diégétique sémantiquement cohérent avec le visuel. La raison tient au fait que dans les autres cas, spatialiser le son reviendrait à créer une scène virtuelle sonore spatialement indépendante de la scène virtuelle visuelle. Or ici, nous nous sommes jusqu'ici intéressés à la coexistence du son et de l'image au sein de la même scène, comme dans notre expérience du chapitre 3. Nous pourrons à ce titre utiliser les résultats et les conclusions que nous en avons tirés, notamment sur le point d'écoute à placer à la position de la caméra.

Enfin, nous considérons le rapport à l'interaction. Notre intérêt pour cette dimension est motivé par le large panel d'interactions possibles sur terminal mobile, qui peuvent potentiellement influencer la qualité d'expérience [Egger, Reichl et Schoenenberg, 2014]. Cette diversité se matérialise par le grand nombre de capteurs embarqués (capteur de pression pour l'écran tactile, microphones, gyroscope, accéléromètre, caméras, capteur de luminosité, récepteur GPS, magnétomètre, thermomètre, etc.), et s'explique sans doute d'une part par l'absence des modes d'interaction classiques du numérique (clavier et souris) et d'autre part par la volonté de s'adapter aux aléas contextuels. [Moller et al., 2009] proposent déjà une taxonomie de l'interaction homme-machine orientée vers la qualité d'expérience. Cependant, les catégories présentées permettent précisément de déterminer si une interaction va aller dans le sens d'une bonne qualité d'expérience ou non, ce qui n'est pas notre but ici. Dans notre

Rapport au contexte	*context-aware*	*context-driven*	passif
Relation entre son et image	son et image corrélés & son diégétique	son et image corrélés & son non-diégétique	son et image non-corrélés
Rapport à l'interaction	pas d'interaction	interaction simple	interaction complexe

TABLE 5.1 – Proposition de taxonomie des applications selon les trois dimensions que nous avons élaborées. En vert, les catégories retenues pour notre choix d'application.

taxonomie, nous cherchons selon quels critères une interaction pourrait influencer la qualité d'expérience du son spatialisé. L'interaction étant une série d'actions et de réactions physiques de l'utilisateur, son aspect gestuel nous intéresse, car il questionne le rapport de l'utilisateur à son propre corps, à l'espace qui l'entoure. Dans cette perspective, nous distinguons les catégories suivantes : absence d'interaction (e.g., application de visionnage de contenu audiovisuel par exemple), interaction simple (utilisant la mobilité tactile uniquement) et interaction complexe (faisant intervenir d'autres modalités d'interaction, la voix, le déplacement physique du corps, etc.). D'autres catégories pourraient être envisagées (relatives par exemple à la durée de l'interaction, ou au type de réponse de l'interlocuteur, qu'il soit humain ou système), mais elles ne seraient pas particulièrement pertinentes vis-à-vis du binaural. Pour en revenir aux catégories retenues, celle de l'interaction simple, avec la modalité tactile, retient notre attention. Elle est pleinement associée au terminal mobile, idiosyncratique pourrions-nous dire, à l'inverse de l'absence d'interaction (qu'on peut retrouver sur tous les terminaux audiovisuels), et sans doute plus répandue auprès du grand public que l'une ou l'autre interaction complexe.

Nous récapitulons les trois dimensions énoncées dans la Table 5.1. Pour chacune d'entre elles, nous avons tenté de fournir des caractéristiques « mutuellement exclusives et collectivement exhaustives », comme le préconisaient [NICKERSON et al., 2009]. Nous avons également discuté pour chaque dimension de la caractéristique que nous avons jugée comme étant la plus pertinente pour notre expérience, la plus représentative de l'utilisation grand public, ou la plus abordable techniquement. Il aurait été intéressant de combiner ces catégories de diverses façons afin d'imaginer plusieurs applications sur lesquelles mesurer l'apport du binaural. Mais les contraintes de ce travail de thèse en temps et en ressources ne nous le permettent pas. Pour cette raison, notre intérêt s'est exclusivement tourné vers le choix suivant : une application utilisable partout, dont le comportement est le même quel que soit le contexte ; audiovisuelle, où le son et l'image sont en relation sémantique directe, avec du son diégétique, c'est-à-dire intégré dans la même scène virtuelle que l'image ; et interactive, via le mode d'interaction principal des terminaux mobiles d'aujourd'hui, à savoir l'écran tactile. La sous-section suivante concrétise ces catégories par un choix d'application.

5.3.2 Choix d'une application

Dans les grandes lignes, notre expérience doit permettre de mesurer l'apport du binaural, c'est-à-dire de comparer une version binaurale de l'application avec une version

non-binaurale (le choix de cette version de référence est discuté dans la section suivante). En plus de la taxonomie, cela implique une contrainte forte sur l'application qui doit donc exister en deux versions. Un type d'applications qui correspond à ces critères est le jeu vidéo, dont le rapprochement avec le son spatialisé se fait de surcroît du plus en plus insistant (en témoignent les nombreuses solutions de spatialisation sonore compatibles avec, voire pensées pour le jeu vidéo, qui émergent depuis quelques années, comme l'Oculus Spatializer de Facebook/Oculus, Resonance Audio de Google ou Steam Audio de Valve[1]). Le jeu vidéo met en scène des sons et des images fabriqués de toutes pièces et agencés artificiellement dans un espace virtuel. Le joueur navigue dans cet espace de façon plus ou moins libre et imprévisible, obligeant à adapter en temps réel le rendu des sources sonores en fonction de son point d'écoute du moment. Ce constat, valable pour un grand nombre de jeux, nécessite l'emploi de la synthèse binaurale, dans une conception de la scène sonore orientée objet, où chaque source est sonorisée indépendamment des autres (conception évoquée par [Rumsey, 2002] notamment, et à plusieurs reprises dans le chapitre 4). L'avantage de la synthèse binaurale pour nous est qu'elle est produite à partir d'un fichier mono d'origine. Par conséquent, grâce à son procédé de production même, la synthèse binaurale nous donne directement accès aux deux versions dont nous avons besoin pour l'expérience.

Plusieurs jeux vidéo accessibles au grand public permettent déjà de vivre une expérience en binaural sur mobile (voir par exemple [Somethin'Else, 2011a ; Somethin'Else, 2011b ; Somethin'Else, 2013 ; DOWINO, 2015]). Néanmoins, ils sont tous purement sonores, ou simplement pourvus d'une interface visuelle de navigation dans l'espace sonore (voir Figure 5.1). À notre connaissance donc, il n'existe pas de jeu vidéo audiovisuel sur mobile en binaural. Il est donc difficile d'envisager de réutiliser un jeu déjà existant. Cette constatation nous pousse à envisager de développer notre propre application. Nous seulement cette solution nous permet d'imaginer le jeu que nous voulons, mais également d'intégrer la collecte des facteurs d'influence et des méthodes d'évaluation directement dans l'application.

La question qui se pose alors est : quel genre de jeu mobile serait le mieux placé pour cette expérience ? Il existe sans doute plusieurs réponses. Nous proposons de nous intéresser au genre de l'*Infinite Runner* (ou *Endless Runner*). Il s'agit d'un type de jeu dans lequel le joueur incarne un personnage qui avance automatiquement, et de plus en plus vite, dans un couloir sans fin. Le but du jeu est d'aller le plus loin possible en esquivant les obstacles et en amassant le plus de bonus possibles. Pour ça, le joueur dispose d'une palette de mouvements qui varie selon les mécaniques du jeu, comme sauter, glisser, se déplacer latéralement, etc. Parmi les titres les plus connus du genre, nous pouvons citer Canabalt [Semi-Secret Software, 2009], Temple Run [Imangi Studios, 2011], Jetpack Joyride [Halfbrick Studios, 2011] ou Subway Surfers [Kiloo & Sybo Games, 2012] (voir Figure 5.2). Ces quatre exemples permettent d'illustrer les deux déclinaisons du genre qu'on peut trouver : le personnage vu de côté en 2D qui progresse de droite à gauche, et le personnage vu de derrière en 3D. C'est plutôt cette seconde version qui nous intéresse, dans la mesure où c'est en 3D, là où les obstacles, objets du décors, etc. passent de part et d'autre de la caméra, que le binaural pourra être utilisé le plus efficacement.

1. http://designingsound.org/2018/03/29/lets-test-3d-audio-spatialization-plugins/

FIGURE 5.1 – Aperçu des interfaces graphiques qui accompagnent le joueur tout au long de son aventure dans trois jeux mobiles en binaural : Papa Sangre (en haut à gauche), Papa Sangre II (en haut à droite) et The Nightjar (en bas).

Les *Infinite Runners* nous intéressent particulièrement pour quatre raisons. Premièrement, c'est un genre de jeu qui est né sur smartphone, conçu pour être en adéquation avec le paradigme de l'interaction tactile (à l'inverse de nombreux genres adaptés d'autres plateformes, console ou PC, qui tentent avec plus ou moins de succès de mimer leur fonctionnement, en émulant par exemple des croix directionnelles ou des touches tactiles). Il offre donc une interaction simple, telle que définie dans notre taxonomie, mais plus encore, une interaction ergonomique. Deuxièmement, les *Infinite Runners* sont particulièrement répandus auprès du grand public (à titre d'exemple, Subway Surfers cumule à lui tout seul plus d'un milliard d'installations rien que sur les plateformes Android[2] !). Si donc ce genre ne représente pas à lui tout seul l'ensemble des jeux sur smartphone, il est tout de même représentatif d'une certaine utilisation grand public, ce qui légitimise son utilisation dans notre expérience. Troisièmement, c'est un genre de jeu qui favorise les parties courtes : le joueur voit ses réflexes mis à rude épreuve, avec une courbe de progression qui mise sur la répétition des parties et l'amélioration par les bonus débloqués, plutôt que sur le développement au cours d'une même partie. De ce fait, c'est un mode de fonctionnement parfaitement adapté à notre protocole expérimentale fondé sur l'ESM. Enfin quatrièmement, le développement de ce genre de jeu est relativement aisé et limité dans la quantité de matériaux graphiques et sonores à produire, car à chaque partie le joueur repart de zéro. Il serait donc envisageable d'en produire un prototype suffisamment avancé pour notre expérience.

Avant de poursuivre cependant, il faut discuter d'un point important concernant la

2. https://play.google.com/store/apps/details?id=com.kiloo.subwaysurf&hl=fr

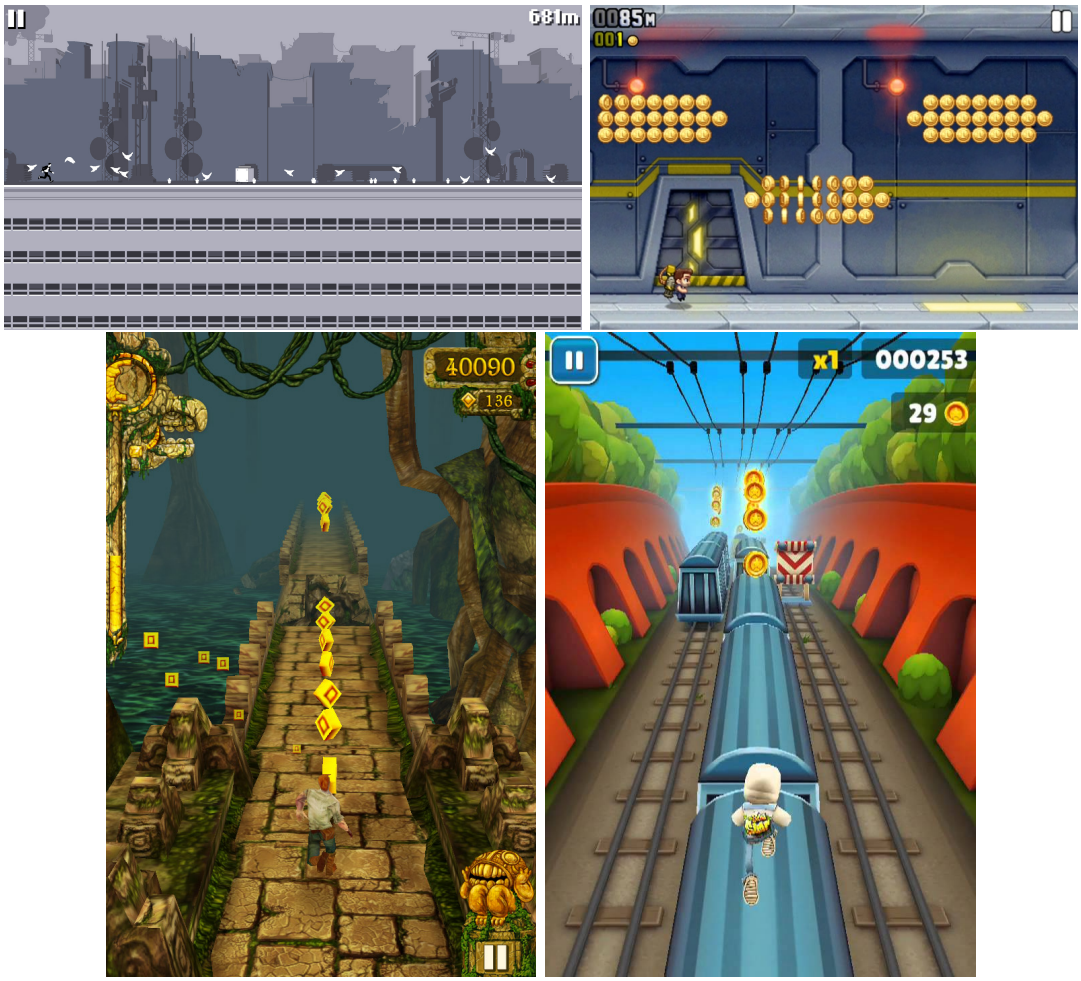


FIGURE 5.2 – Captures d'écran de quatre jeux vidéo de type *Infinite Runner* : Canabalt (en haut à gauche), Jetpack Joyride (en haut à droite), Temple Run (en bas à gauche) et Subway Surfers (en bas à droite).

validité des données. Nous avons adopté l'ESM à la faveur de la validité externe des données, mais qui soumet l'expérience à l'influence de facteurs contextuels, impliquant une baisse de validité interne. Un autre aspect de la validité interne des données concerne la répétabilité des parties. Dans un *Infinite Runner*, les niveaux sont générés procéduralement, c'est-à-dire que les obstacles, les ennemis, les bonus, les embranchements, les virages, tout est en partie ou en totalité sélectionnés et placés sur la route du joueur dans un ordre aléatoire (ou pseudo aléatoire). En principe donc, deux parties ont très peu de chance d'être identiques. Cela se traduit dans notre expérience par le fait que deux sujets, sur le même numéro de session, vont nécessairement accomplir deux parties différentes, dont une sera potentiellement plus difficile que l'autre, avec plus d'obstacles, un environnement moins apprécié du sujet, etc., amenant une source de biais dans les données. Cependant, contrôler ces variations serait extrêmement difficile, posant des problèmes de conception de niveau pour ainsi dire insolubles. D'abord, puisque la partie se poursuit tant que le joueur ne meurt pas, il est impossible de s'assurer que deux parties sont de même durée. La limiter artificiellement reviendrait à modifier radicalement les mécaniques du jeu (le jeu ne serait plus un *Infinite Runner* par définition), affectant en cela le ressenti du joueur et la validité externe des données. Nous pourrions envisager des solutions d'atténuation des biais en concevant par exemple des niveaux finis qui bouclent sur eux-mêmes, afin que le joueur reste dans un cadre contrôlé. Mais là aussi, cela pose problème. Il faudrait d'une part que chaque niveau soit suffisamment grand pour que le joueur n'ait pas l'impression de tourner en rond, et d'autre part produire autant de ces niveaux qu'il y a de sessions de jeu. Non seulement le travail de conception serait quantitativement faramineux (allant bien au-delà du délai permis dans cette thèse), mais aussi qualitativement très difficile : il faudrait s'assurer que chaque niveau propose une difficulté progressive et équilibrée, avec un placement de tous les éléments du jeu mûrement réfléchi. Dans une telle étape de conception, les studios de développement de jeux vidéo ont coutume de faire appel à de nombreux testeurs, dans une démarche d'allers-retours longue et coûteuse en ressources que nous ne pouvons pas nous permettre. La solution que nous proposons donc est de s'en remettre au hasard, c'est-à-dire à la variabilité des niveaux générés aléatoirement. Nous faisons donc l'hypothèse que, sur la totalité des sessions, chaque joueur aura rencontré en moyenne les mêmes difficultés que les autres. Nous considérons cette hypothèse raisonnable dans la mesure où c'est celle qui est adoptée en pratique dans les *Infinite Runner* publiés sur le marché.

Le jeu développé pour l'expérience sera présenté en détails dans le chapitre 6. Il nous faut avant aborder le dernier point de la problématique, sur les types d'apports possibles du binaural, ceux que nous retenons, et la façon dont ils peuvent se mesurer dans un *Infinite Runner*.

5.4 Quel apport du binaural mesure-t-on ?

5.4.1 Le choix d'une version de référence avec laquelle comparer le binaural

Nous voulons évaluer l'apport du binaural dans un jeu vidéo en comparant une version binaurale de ce jeu à une autre version de référence, sonorisée différemment. Nous

voyons trois possibilités pour cette référence : une version muette, une version monophonique ou une version stéréophonique. Nous nous contentons d'une comparaison avec un seul système de référence pour éviter de surcharger l'expérience. À première vue, les trois pourraient prétendre à une comparaison légitime avec le binaural : la stéréo est le mode de restitution sonore sur casque grand public le plus élaboré du point de vue de la spatialisation ; la mono est le mode de restitution privilégié (voire exclusif) dans les *Infinite Runners* présents sur le marché ; tandis que l'absence de son est sans doute un mode d'utilisation largement répandu auprès du grand public. En revanche, si l'on ne considère la comparaison qu'en termes purement acoustiques, une hiérarchie s'installe : un son stéréo permettrait de mettre en avant l'apport propre du binaural, tandis qu'un son mono ne permettrait de mettre en avant que l'apport de la spatialisation dans son ensemble, et une version muette ne permettrait de conclure que sur l'intérêt ou non de la présence de son, quel qu'il soit, indépendamment de son niveau de spatialisation. Dans l'idéal donc, pour mesurer l'apport du binaural en tant que tel, il faudrait le comparer avec de la stéréo. Toutefois, un autre argument nous fait pencher en faveur de la monophonie. De par l'utilisation de la méthode ESM, nous savons que nous allons mesurer l'apport du binaural dans des contextes potentiellement perturbants, aussi variés qu'incontrôlables. L'objectif de cette expérience sera donc de voir si, malgré tout le bruit induit, un effet significatif du binaural peut tout de même être observé. L'ESM étant encore relativement peu usitée dans l'état de l'art, il nous semble important, dans la perspective de valider l'aspect méthodologique de cette expérience, d'ajuster les quelques paramètres sur lesquels nous avons le contrôle de façon à faire émerger au mieux des informations saillantes. En cela, la monophonie comme version de référence nous semble un bon compromis : elle est suffisamment différente du binaural pour permettre une comparaison malgré les contextes bruyants, et permet en même temps de mettre en avant les bénéfices de la spatialisation sonore.

5.4.2 Des attributs pour évaluer le binaural

Il nous faut maintenant choisir selon quels critères nous allons comparer ces deux versions. Dans la section 4.3.2 du chapitre 4, nous avons présenté de nombreux attributs selon lesquels le son spatialisé serait à même d'être évalué. Toutefois, il serait sans doute peu pertinent ici de mesurer l'apport du binaural dans des termes purement auditifs. En effet, notre but n'est pas tant d'accéder directement à la qualité sonore du binaural que de mesurer son apport à la qualité d'expérience générale.

De ce fait, il nous faut choisir des attributs d'expérience plus généraux que ceux de la qualité sonore, mais en même temps des attributs susceptibles d'être affectés par le son. Parmi tous les attributs passés en revue dans le Chapitre 4, trois retiennent notre attention : le sentiment d'immersion, la mémorisation et la performance. Ils nous intéressent particulièrement car non seulement ils font partie de ces attributs évoqués par l'état de l'art qui ne sont pas circonscrits au domaine du son, mais aussi parce qu'ils sont évalués via des méthodes de mesure différentes : l'immersion par une méthode subjective (un questionnaire), la mémorisation par l'accomplissement d'une tâche après la session (méthode objective), et la performance par une récupération des données d'interaction pendant la session (autre méthode objective). Dans notre expérience où nous privilégions la validité externe des données à leur validité interne, il peut être judicieux de multiplier les méthodes d'évaluation, une façon pour nous de

recouper les informations et de compenser les biais intempestifs qu'une méthode peut subir à cause de tel ou tel facteur d'influence.

Immersion et mémorisation sont toutes les deux évaluées en sollicitant le sujet. Pour éviter un effet d'anticipation du sujet sur les questions à venir, nous envisageons d'alterner entre questionnaire sur l'immersion, tâche de mémorisation et absence de tâche. Cette répartition des tâches impliquera un nombre moindre d'informations recueillies à chaque session, mais permettra, en plus de gommer l'effet d'anticipation, de raccourcir la tâche post-session, pour coller au mieux avec la courte durée imposée par l'ESM.

Immersion

L'immersion (et les notions associées comme le sentiment de présence ou l'engagement) fait partie de ces attributs largement identifiés à la fois dans le domaine du jeu vidéo (voir par exemple l'état de l'art de [Soriano, 2016]) et dans celui du son spatialisé [Berg et Rumsey, 2002 ; Hamasaki, Nishiguchi et al., 2004 ; Guastavino et Katz, 2004 ; Blauert et Jekosch, 2012 ; Nicol, Gros, Colomes, Roncière et al., 2016 ; Simon, Zacharov et Katz, 2016 ; Katz et Nicol, 2018]. En particulier, des études se sont attachées à mettre au point des questionnaires permettant de la mesurer au mieux, à la fois dans le jeu vidéo [Brockmyer et al., 2009], pour l'expérience d'environnements virtuels [Witmer et Singer, 1998] et pour évaluer du son spatialisé [Nicol, Gros, Colomes, Roncière et al., 2016]. Nous nous en inspirons pour mettre au point nos questions, toujours avec le souci de limiter la durée des sessions et donc des questionnaires :

1. « Dans le monde généré par le jeu, vous avez eu le sentiment "d'y être" » : par cette assertion, nous questionnons davantage le sentiment de présence, plutôt que l'immersion, car c'est l'attribut qui semble être le plus revenu dans les études sur le son spatialisé.
2. « Avez-vous ressenti un effet "son 3D" ? » : on s'assure ici que le son binaural est bien détecté comme tel. Par ailleurs, le sentiment de présence induit par un jeu vidéo avec des graphismes en 3D étant potentiellement fort, il pourrait favoriser une perception de son 3D même quand il n'y en a pas. De ce fait, cette question est aussi une façon de s'assurer que le son monophonique est aussi bien perçu comme tel. Avec cette question nous avons quatre réponses possibles : son 3D détecté dans les sessions binaurales (vrai positif), son 3D non-détecté dans les sessions mono (vrai négatif), son 3D détecté dans les sessions mono (faux positif), son 3D non-détecté dans les sessions binaurales (faux négatif).
3. « Le son a contribué à votre sentiment d'immersion » : cette assertion permet de savoir dans quelle mesure le son participe au sentiment général d'immersion. Ici aussi, l'idée est de comparer binaural et monophonie, savoir si le premier contribue davantage au sentiment d'immersion que le second, et dans quelle mesure.
4. « Le contexte extérieur (bruit ambiant, tâches menées en parallèle, etc.) a gêné votre immersion » : cette question est à recouper d'une part avec les questions contextuelles, pour voir quels sont les contextes les plus perturbants. Mais surtout, elle nous permet de mesurer le pouvoir isolant du binaural par rapport à la monophonie.

À l'exception de la deuxième question, qui offrira un simple « Oui » ou « Non » comme choix de réponse, les réponses seront fournies sur un échelle de notation numérique, discrète, à quatre degrés, dont les extrémités seront annotées par « Tout à fait » et« Pas du tout ». Ce choix d'un nombre pair d'échelons nous permet de contraindre le sujet à une réponse tranchée (un nombre impair autorisant à l'inverse une position centrale neutre), alors que des contextes d'utilisation perturbants l'auraient peut-être incité à une prudence pas nécessairement représentative de sa perception.

Mémorisation

La mémorisation est un autre de ces attributs susceptibles d'être améliorés par la spatialisation sonore. La relation entre mémoire et espace existe depuis l'Antiquité [Yates et Arasse, 1987]. Elle émane de la volonté de l'être humain à organiser sa pensée au moyen de l'imagination. L'art de la mémoire est une technique qui consiste à se représenter mentalement un lieu dans lequel on place des images marquantes, qui servent à se souvenir d'éléments précis (des choses ou des mots, selon la catégorisation primitive). Une promenade mentale dans ce lieu aidera à convoquer ces images dans un ordre précis, et à se remémorer les éléments associés. Si cette technique de mémorisation a traversé l'histoire sous diverses formes pour nous atteindre aujourd'hui (utilisée notamment par les champions du monde de la mémoire[3]), elle a également fait l'objet d'études intéressantes sur sa relation à l'informatique. Dans sa thèse récente, [Aubert, 2019] montre que, dès le milieu du XX^{e} siècle, une passerelle était établie par certains entre espace mental et cyberespace, ce dernier étant vu comme une opportunité d'étendre les capacités mémorielles de l'être humain via les nouveaux outils de représentation et de stockage de données. L'arrivée des espaces virtuels retransmis en audiovisuel a encore renforcé cette analogie, déjà existante par ailleurs au cinéma. Dès le début des années 90, des jeux vidéo, comme par exemple le très populaire Myst [Cyan World, 1993], exigeaient une connaissance parfaite de la géographie des lieux dans lesquels on se déplaçait pour résoudre des énigmes, nécessitant de se constituer un double intérieur de cet espace, comme le commande précisément l'antique art de la mémoire.

Notre postulat est que le binaural peut être un moyen intéressant d'approfondir le lien entre scène virtuelle et scène mentale, en facilitant l'impression de spatialité chez le joueur. Dans l'expérience de [Larsson, Vastfjall et Kleiner, 2002], la mémorisation a été montrée comme étant meilleure pour des stimuli avec son que pour les stimuli sans son, mais aucune différence significative n'a été révélée entre la version stéréo et la version binaurale. Cependant, comme nous l'avons déjà exposé, pour connaître l'apport de la spatialisation sonore, nous choisissons ici de comparer binaural et monophonie.

L'art de la mémoire traditionnelle recommande de poster des images marquantes le long d'un chemin qu'on est amené à parcourir autant de fois que nécessaire pour les imprimer dans le souvenir. Dans un *Infinite Runner*, la scène virtuelle est effectivement composée d'un chemin, mais celui-ci diffère systématiquement d'une partie à l'autre, puisqu'il est généré aléatoirement. Il ne nous est donc pas possible de mesurer le résultat d'un apprentissage mnémotechnique sur plusieurs sessions. En revanche,

3. Voir par exemple le site `http://www.artdelamemoire.org/`, tenu par l'ex-champion du Canada

nous pouvons mesurer la mémoire à court terme, pour laquelle l'art de la mémoire a également fait ses preuves. Nous proposons donc la tâche de mémorisation suivante : pendant sa session de jeu, le joueur croise différents objets audiovisuels qui jalonnent le bord du chemin. À la fin de la session, il doit se rappeler l'ordre des objets qu'il aura croisés. Comme on ne connait pas à l'avance la durée de sa session, nous ne l'interrogeons que sur les sept derniers objets croisés (l'empan mnésique de la mémoire à court terme, en dehors de toute technique de mémorisation, ayant été identifiée comme avoisinant le chiffre sept [Miller, 1956]).

Une attention particulière devra être portée sur le choix des objets : les traités sur la mémoire précisent bien que seules des images marquantes peuvent facilement se rappeler à la mémoire. Cette précision impliquera sans doute des contraintes sur la conception graphique : des objets qui dénotent de l'environnement, soit par leur type, soit par leur forme, et les répercussions sur le rendu général.

Pour évaluer la mémorisation du sujet, la tâche de restitution de l'ordre des objets devra suivre la session. On présentera une liste des sept derniers objets croisés dans le désordre ; le sujet devra les remettre dans le bon ordre. En supplément, nous souhaitons avoir une idée du rôle que le son a joué dans cette tâche (toujours pour comparer bien sûr entre le binaural et la monophonie). Pour cette raison, nous ajoutons deux assertions que le sujet devra noter : « Le positionnement sonore de ces objets vous a aidé à les classer » et « Le positionnement visuel de ces objets vous a aidé à les classer », accompagnées de la même échelle à quatre degrés que pour l'immersion.

Performance

La performance est un attribut très couramment évalué dans les jeux vidéo au moyen d'un score. C'est une mécanique qui concrétise une progression ou un accomplissement du joueur, souvent utilisée pour des jeux ou parties de jeux sans jalon scénaristique. Dans les *Infinite Runners*, le score combine la distance parcourue et les bonus amassés. D'une certaine façon, il symbolise la capacité du joueur à avoir su se frayer un chemin à travers les obstacles, et s'apparente de ce fait aux tâches de navigation qu'on utilise régulièrement pour évaluer des scènes sonores spatialisées [Larsson, Vastfjall et Kleiner, 2002 ; Gonot, Emerit et Château, 2006]. Nous relèverons donc cette information naturellement présente dans le jeu, pour comparer les scores obtenus en mono avec ceux du binaural.

Remarques sur les attributs et la qualité d'expérience

Quelles relations entretiennent ces trois attributs ? C'est une question que nous laissons délibérément de côté, et sur laquelle nous reviendrons ou non en fonction des résultats de l'expérience. Par conséquent, nous ne nous préoccupons pas ici de leur orthogonalité. Il est possible qu'ils soient indépendants, ou liés entre eux d'une manière ou d'une autre (par exemple [Aubert, 2019] relie informellement la mémorisation spatialisée au sentiment de présence). Comme nous l'avons déjà dit, la diversité des méthodes d'évaluation a davantage guidé notre réflexion sur le choix des attributs.

Par ailleurs, est-ce qu'en mesurant l'apport du binaural en termes d'immersion, de mémorisation et de performance, on mesure un apport du binaural en termes de qualité d'expérience ? Ce n'est pas sûr. De par la complexité des rapports entre les attributs, une amélioration de l'un ou l'autre pourrait très bien être contrebalancée par une baisse ailleurs, débouchant sur une qualité globale neutre. Par exemple, une hausse de performance, mais une baisse de mémorisation. Nous citions dans la section 4.4.3 du chapitre 4 l'expérience de [Jennett et al., 2008], dans laquelle les sujets jouent à un jeu et ressentent une forte immersion en même temps qu'un affect émotionnel négatif. Difficile de conclure dans ces conditions sur une valeur de qualité d'expérience générale. Étant donné le peu de place que nous laisse la méthode ESM pour solliciter le sujet à chaque session, nous préférons nous concentrer sur nos attributs, plutôt que de réserver une place supplémentaire à une évaluation globale. À notre charge donc d'analyser et d'interpréter les résultats à l'issue de l'expérience pour conclure sur l'apport global du binaural.

5.5 En résumé, une expérience entre bien-fondé théorique et faisabilité pratique

Ce chapitre a été l'occasion d'exposer les différents aspects de la problématique de l'évaluation de l'apport du binaural dans une application mobile audiovisuelle. L'expérience que nous présentons peut se résumer de la façon suivante : une mesure de l'apport du binaural en termes d'immersion, de mémorisation et de performance sur un jeu vidéo mobile de type *Infinite Runner*, menée grâce à la méthode ESM qui fractionne l'expérience en une multitude de petites sessions, replacées dans le contexte d'utilisation quotidien des sujets. Les réflexions exposées dans ce chapitre nous ont permis de situer nos choix au sein d'un état de l'art qui, nous l'espérons, leur confère une certaine validité théorique. Malgré tout, nous gardons à l'esprit que certaines restrictions d'ordre pratique (une seule application utilisée, seulement trois attributs évalués), en même temps que des absences de restriction (contexte d'expérimentation écologique, donc non contrôlé), donnent à cette expérience et à ses résultats une portée limitée. D'une certaine façon, l'orientation « validité externe » qui nous a guidés dans ce chapitre, et qui va dans le même sens que d'autres études récentes, accorde à cette étude un caractère méthodologique exploratoire qui nous semble faire pleinement partie de la problématique de cette thèse, tout autant que l'apport du binaural en lui-même.

Dans le chapitre suivant, nous présentons l'expérience qui découle de ces choix, en commençant par le développement de l'application, le déroulement de l'expérience avec le passage des sujets, l'observation, l'analyse et l'interprétation des résultats. Une attention sera portée non seulement sur les données en elles-mêmes, mais également sur les détails de déploiement de l'expérience, leurs avantages et leurs inconvénients.

Chapitre 6

Mesure de l'apport du binaural dans un *Infinite Runner*

6.1 Introduction

Dans ce chapitre nous présentons l'expérience qui fait suite aux choix du chapitre précédent. La section 6.2 commence par présenter le jeu qui a été développé pour l'occasion. La section 6.3 présente le protocole, la section 6.4 les résultats et la section 6.5 une discussion sur l'expérience. Cette dernière permettra non seulement d'interpréter les résultats, mais de discuter également de la méthode de déploiement, l'*Experience Sampling Method* (ESM). La section 6.6 concluera ce chapitre.

6.2 Le jeu vidéo

Le jeu vidéo est conçu avec l'aide de prestataires extérieurs : Polymorph Studio [1] pour le développement et Studio Anatole [2] pour l'intégration audio. Le jeu est intégralement développé avec le moteur de jeu vidéo Unity. En guise de cahier des charges fonctionnel, un prototype est développé par nos soins (jusqu'à un stade trop peu avancé pour constituer un jeu définitif) pour définir les principales mécaniques de jeu que nous souhaitons retrouver dans l'*Infinite Runner* : un jeu en 3D, avec une caméra placée derrière l'avatar du joueur ; un chemin séparé en trois parties – couloirs de gauche, central et de droite – ; avec des bonus ou des obstacles distribués aléatoirement sur ces parties ; la possibilité pour le joueur de se déplacer latéralement sur une partie, de sauter par dessus ou de glisser en-dessous d'obstacles, et enfin surtout la présence de virages, pour créer des changements d'orientation spatiale et favoriser des effets sonores spatialisés. Quelques-uns de ces prérequis sont illustrés dans la Figure 6.1.

Dans la version finale, le joueur possède trois vies, lui permettant de perdre sans mettre fin immédiatement à la session. Le jeu possède par ailleurs un affichage tête haute (*Head-up display*, ou HUD), une interface graphique superposée à l'écran de jeu qui affiche diverses informations (nombre de vies restantes, score actuel, etc.)

1. https://www.polymorph.fr/
2. http://studio-anatole.com/

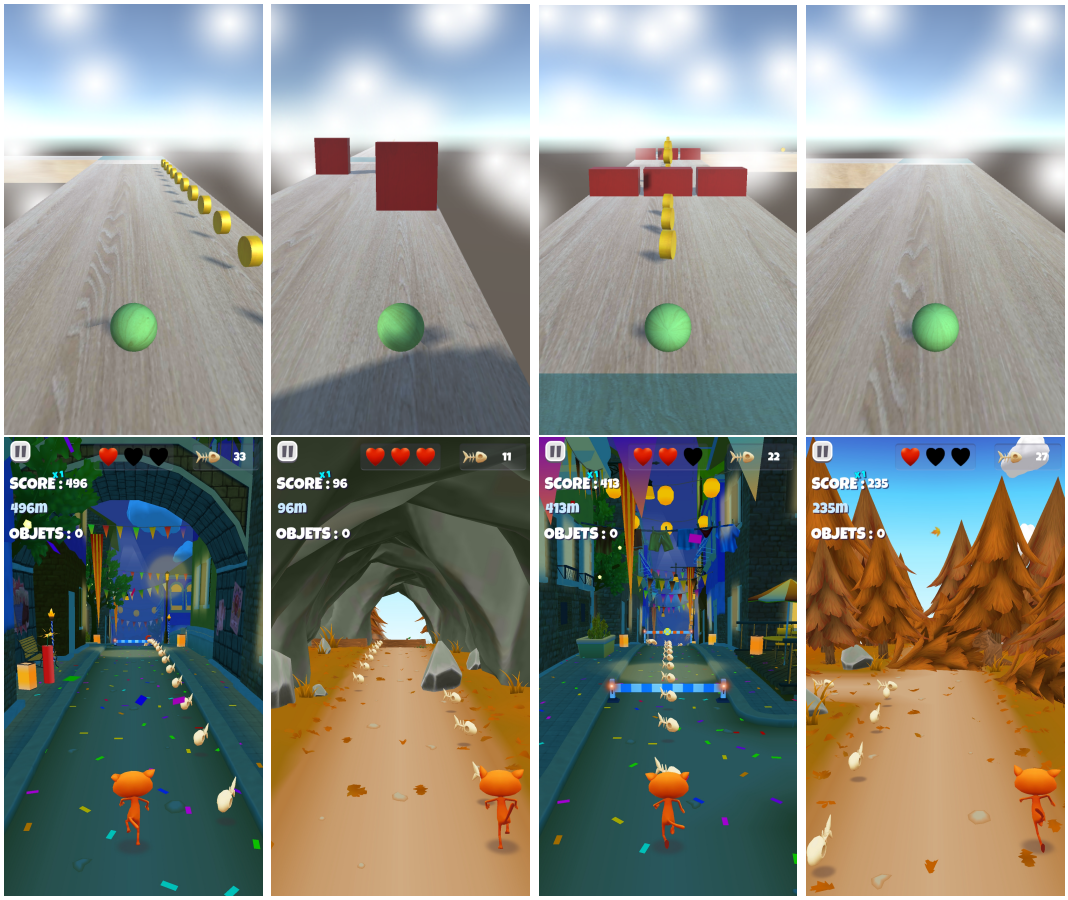


FIGURE 6.1 – Illustrations du prototype au-dessus, et de la version finale en dessous. De gauche à droite on peut observer : les collectables, disséminés ici sur la partie droite du chemin ; un obstacle à éviter ; un obstacle par dessus lequel sauter ; et un virage à venir, qui modifie l'orientation spatiale du joueur et donc sa perception des sons.

Pour le contenu graphique, le prestataire s'est appuyé sur des ressources gratuites disponibles sur la boutique en ligne de Unity, l'Asset Store [3]. Le style graphique *cartoon* (voir Figure 6.1), correspond bien à l'exigence de placer des images de mémoire qui dénotent sémantiquement de l'environnement, sans pour autant altérer la crédibilité de la scène virtuelle (le cartoon étant propice à ce genre de fantaisie). Vingt-trois objets de mémoire sont modélisés, répertoriés dans la Figure 6.2. En plus de leur caractère sémantiquement inattendu, ils sont aussi choisis pour leurs sons marqués et facilement identifiables.

Pour apporter de la diversité au jeu, trois décors sont implémentés (environnements forestier, urbain et vénitien), chacun décliné en deux versions (forêt printanière ou automnale, ville normale ou enneigée, ville vénitienne normale ou en carnaval), et deux modes (jour ou nuit). L'intérêt d'une telle variété est de minimiser la lassitude du joueur au fil de l'expérience.

Par ailleurs, le son binaural est intégré à l'aide d'un plugin Unity, développé en interne par Marc Émerit (chercheur à Orange Labs), autorisant la synthèse binaurale en temps-réel avec des HRTF individualisées ou non au format SOFA [Majdak et al., 2013]. L'intégration sonore en binaural suit nos recommandations issues de l'expérience présentée au chapitre 3 : point d'écoute placé sur la caméra et HRTF non-individualisées (issues de la base de données de HRTF d'Orange, mesurées sur une tête artificielle KU100 de Neumann).

Les sons binauraux sont de deux types : des sources ponctuelles associées à des objets précis (obstacles, objets de mémoire, bonus), ou des sons d'ambiance, sources ambisoniques binauralisées, pour retranscrire l'environnement plus lointain. Tous ces sons existent aussi en version mono. Par ailleurs, tous les sons d'interface sont rendus uniquement en mono, car non-diégétiques.

6.3 Protocole

6.3.1 Déroulement global de l'expérience, plan des sessions

L'expérience dure au total pour chaque sujet 5 semaines, pour 70 sessions, à raison de 2 sessions par jour. Afin de diversifier les contextes d'utilisation, une session a lieu le matin, l'autre l'après-midi, à des heures aléatoires respectivement comprises entre 8h et 13h et 13h et 18h. Le sujet n'est pas informé à l'avance de ses heures de session. Il reçoit une notification par SMS de l'expérimentateur le moment venu, et a pour instruction d'accomplir sa session dès que possible. S'il rate une session, l'ensemble de la journée est reportée, rallongeant la durée totale de l'expérience d'un jour (si le sujet rate le matin, le matin et l'après-midi sont reportés ; si le sujet rate l'après-midi, seul l'après-midi est reporté, mais le lendemain ne comporte pas de session le matin). De cette façon, les sessions restent réparties équitablement entre matin et après-midi.

3. https://assetstore.unity.com/packages/essentials/tutorial-projects/endless-runner-sample-game-87901

FIGURE 6.2 – Modèles 3D des objets de mémorisation. Avec, de gauche à droite et de haut en bas : un aspirateur, une cloche, une corne de brume, une crécelle, un ghetto-blaster, un gramophone, une guitare, un klaxon, une machine à laver, une maraca, un marteau-piqueur, un mixeur, un pétard, un piano, un réveil, un robot, une roulette de dentiste, un sèche-cheveux, un tambour, un téléphone, une tondeuse à gazon, une trompette et une tronçonneuse.

Sur les 70 sessions, 35 sont en binaural et 35 en mono. Par ailleurs, sur les 70 sessions, 30 proposent un questionnaire relatif à l'immersion, 30 demandent au sujet d'accomplir la tâche de mémorisation, et 10 sont dépourvues de questionnaire de fin. En combinant type de rendu sonore, type de questionnaire de fin, et moment de la journée, on obtient la répartition suivante :

— 7 sessions en binaural le matin avec questions sur l'immersion ;
— 8 sessions en binaural l'après-midi avec questions sur l'immersion ;
— 8 sessions en binaural le matin avec questions sur la mémorisation ;
— 7 sessions en binaural l'après-midi avec questions sur la mémorisation ;
— 2 sessions en binaural le matin sans question ;
— 3 sessions en binaural l'après-midi sans question ;
— 8 sessions en mono le matin avec questions sur l'immersion ;
— 7 sessions en mono l'après-midi avec questions sur l'immersion ;
— 7 sessions en mono le matin avec questions sur la mémorisation ;
— 8 sessions en mono l'après-midi avec questions sur la mémorisation ;
— 3 sessions en mono le matin sans question ;
— 2 sessions en mono l'après-midi sans question.

Ces sessions sont présentées dans un ordre aléatoire différent pour chaque sujet.

Lorsque le sujet a accompli une session, le bouton de lancement du jeu se grise et devient inaccessible jusqu'à la tranche horaire de la session suivante. Si malgré tout le sujet souhaite jouer davantage, un mode de jeu appelé « Session libre » est disponible sur la page d'accueil, parfaitement identique à l'autre mode (appelé par opposition « Session de test »), mais qui ne fait l'objet d'aucune restriction temporelle. Pour ce mode, le type de spatialisation est choisi aléatoirement, ainsi que la présence ou non d'un questionnaire de fin. Les données des sessions libres sont également récoltées.

Pour entretenir la motivation des sujets au cours de l'expérience, les notifications SMS annoncent progressivement l'argent gagné en cours d'expérience (uniquement pour les sujets payés, voir section 6.3.3 à ce sujet). Au bout des 20 premières sessions, le SMS annonce que le sujet a déjà gagné 10 euros grâce à son assiduité. Au bout de 40 sessions, le message annonce 20 euros gagnés au total. Enfin, à l'issue des 70 sessions, 50 euros reviennent au sujet.

6.3.2 Déroulement d'une session

Une session se déroule en quatre phases : questionnaire de contexte, calibration sonore, phase de jeu et questionnaire de fin. Le questionnaire sur le contexte a été détaillé dans le chapitre précédent, section 5.2.2. La Figure 6.3 montre son intégration dans l'application.

Après le contexte, un panneau sur la calibration sonore arrive. Cette étape sert à contrôler le volume sonore utilisé pendant le jeu. On rappelle d'abord au sujet de bien brancher son casque, puis on lui demande de régler le volume du téléphone au maximum. Enfin, un son est joué pendant lequel le sujet peut ajuster un curseur pour

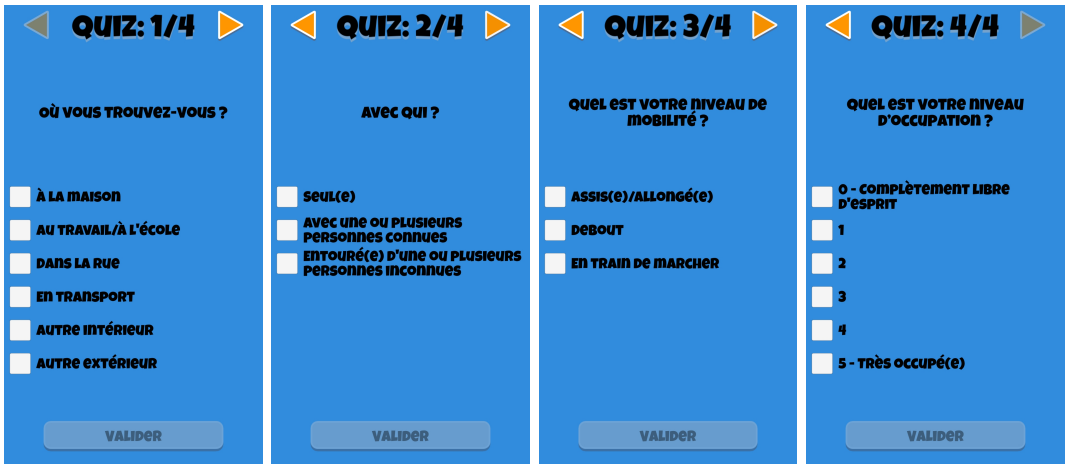


FIGURE 6.3 – Le questionnaire contextuel intégré dans l'application.

régler le volume sonore de l'application selon son confort. Nous récupérons la valeur de ce curseur à chaque session.

La phase de jeu est bien sûr l'étape principale de la session. En plus de la tâche principale – parcourir la plus grande distance possible –, le joueur croise des objets de mémorisation audiovisuels disséminés aléatoirement sur le bord du chemin (voir Figure 6.4 pour un exemple). Comme le joueur est interrogé sur les 7 derniers croisés (voir section 5.4.2 du chapitre précédent sur ce chiffre), il doit en avoir croisé au moins 7. Dans le cas où il échouerait avant, un texte s'affiche expliquant que la session ne peut pas encore être validée, puis le jeu redémarre au même point, les trois cœurs remplis à nouveau (à noter qu'en session libre, le sujet n'a pas cette restriction des 7 objets, et ne se voit par conséquent jamais proposer de questionnaire de mémorisation dans la dernière phase).

Enfin, la dernière phase est celle des questionnaires. De même que pour le contexte, les questions d'immersion et de mémorisation ont aussi déjà été détaillées dans le chapitre précédent, section 5.4.2. Les Figures 6.5 et 6.6 illustrent la façon dont ces questions sont intégrées dans le jeu.

Toutes les données du jeu sont envoyées à l'issue de la session sur un serveur distant du prestataire Polymorph, qui nous a gracieusement octroyé une base de données avec un accès total durant notre expérience. Les sujets doivent donc avoir accès à un réseau mobile ou au Wifi à chaque session. Dans le cas contraire, les données sont sauvegardées en local sur leur téléphone et envoyées à leur prochaine session.

6.3.3 Participants et accueil

Trente sujets participent à l'expérience (de 16 à 67 ans, moyenne de 28,1 ans, 8 femmes). Parmi eux, cinq travaillent à Orange et ne peuvent pas être défrayés. Les vingt-cinq autres sont recrutés à l'extérieur et défrayés par un bon d'achat de 50 euros. Comme l'expérience se déroule avec le matériel des sujets, la campagne de recrutement

FIGURE 6.4 – Ici entouré en rouge, un aspirateur, exemple d'objet de mémorisation qu'on peut croiser dans le jeu. Pour être plus visibles dans le jeu, les objets tournent sur eux-mêmes et grossissent-rétrécissent périodiquement.

FIGURE 6.5 – Le questionnaire d'immersion intégré dans l'application.

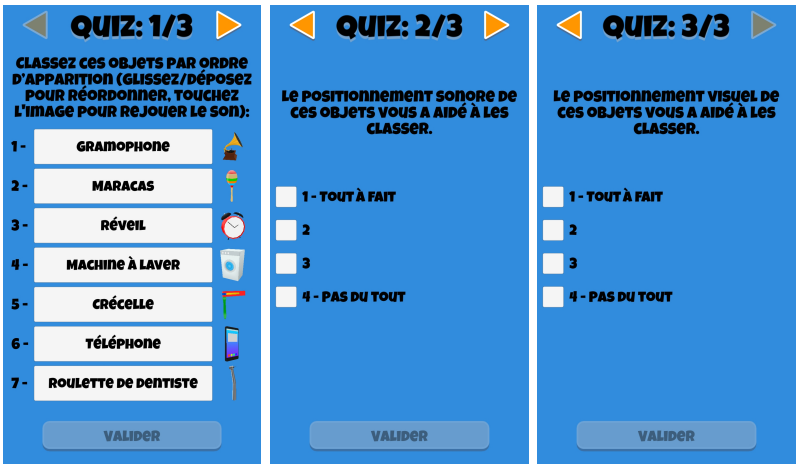


FIGURE 6.6 – Le questionnaire de mémorisation intégré dans l'application.

spécifie comme prérequis la possession d'un smartphone Android (le jeu ne fonctionne que sous ce système d'exploitation) et d'un casque audio, à l'exclusion d'écouteurs intra-auriculaires, qui peuvent bouger en cours d'utilisation. Cette dernière exigence nous assure un minimum de stabilité de la scène binaurale au cours d'une même session. Par ailleurs, il est demandé aux sujets d'avoir une expérience minimum de jeu sur terminal mobile.

L'expérience commence par une rencontre en laboratoire (pour 5 sujets cette phase préliminaire est faite à distance par visioconférence). À leur accueil par l'expérimentateur, 20 sujets déclarent jouer plusieurs fois par semaine sur terminal mobile, 7 plusieurs fois par mois, 1 moins d'une fois par mois et 2 jamais. Ces trois derniers ont néanmoins déjà joué à des jeux mobiles, et connaissent en principe le fonctionnement d'un *Infinite Runner*. D'autres informations sont récoltées, comme le modèle du téléphone, la version d'Android et le modèle du casque, reportées ici dans le tableau 6.1. On constate qu'à l'exception du sujet 13, toutes les versions d'Android sont supérieures ou égales à la 6.0.0, assurant une certaine stabilité du jeu. Plusieurs essais infructueux ont été menés sur des versions plus anciennes, voir la section 6.5.2 pour une discussion sur ce sujet.

Avant de prendre part à l'expérience, chaque sujet est invité à signer une décharge autorisant l'installation de l'application sur son téléphone et détaillant le type des données collectées. L'application est ensuite installée et son bon fonctionnement vérifié (notamment la communication avec le serveur).

Une feuille d'instruction est remise au sujet, avec un résumé sur le déroulement des opérations (nombre total de sessions, rémunération le cas échéant, déroulement d'une session, retard et rattrapage des sessions, etc. Voir le document complet en annexe B.1). Une session libre est effectuée en présence de l'expérimentateur pour vérifier que chaque étape est bien comprise. L'expérimentateur s'assure par ailleurs que le sujet comprend la signification du terme « son 3D », présent dans les questions d'immersion. Cette précision, en plus du port obligatoire du casque, révèle au sujet l'importance du son dans cette expérience. Toutefois, rien ne lui est dit sur la présence des deux types de rendus sonores, ni sur le fait que du son 3D est effectivement utilisé pendant

Sujet	Modèle de téléphone	Version d'Android	Modèle de casque
1	Samsung Galaxy S8	8.0.0	Sony MDR XB450
2	Sony Xperia XZ	8.1.0	Sony MDR ZX660
3	Huawei P8	6.0.0	Sennheiser MM400X
4	Samsung Galaxy A5	6.0.1	Silvercrest
5	Samsung Galaxy A5 2017	8.0.0	Sony MDR ZX770
6	Huawei Ale L21	6.0.0	Silvercrest
7	Samsung Galaxy S8	8.0.0	Philips SHL3665
8	Samsung Galaxy S8	8.0.0	Sony MDR ZX310
9	Meizu MX-5	6.0.0	Gamecom
10	One Plus 6	8.1.0	Sony MDR ZX310
11	Honor View 10	8.0.0	Sony MDR ZX110
12	Samsung Galaxy J7	7.0.0	Philips
13	Samsung Galaxy Tab2	5.1.1	SuperLux HD681
14	Samsung Galaxy A5	7.0.0	Sennheiser HD280 Pro
15	Samsung Galaxy S9+	8.0.0	AKG NG60 NC
16	Honor 7X	8.0.0	JBL T450BT
17	Honor 9	8.0.0	Sony MDR ZX310
18	Samsung Galaxy S9	8.0.0	AKG K845
19	Homtom HT20Pro	6.0.0	Sony MRX 1000
20	Alcatel Idol4	6.0.1	-
21	Thor E	7.0.0	Razer Kraken
22	Asus Zenfone 3 Max	7.0.0	Razer
23	Asus Zenfone 4 Max Plus	7.1.1	Beats By Dre
24	Samsung Galaxy S6 Edge	7.0.0	Sog
25	Huawei P8 Lite 2017	8.0.0	New One HD65
26	Samsung Galaxy S5	7.0.0	Grundig
27	Vivo Y55A	6.0.1	Sennheiser
28	Samsung Galaxy J3	7.0.0	Gamecom 388 Plantronics
29	LG G5 SE	7.0.0	AKG Y50 BT
30	Honor 9 Lite	8.0.0	Audio Technica ATH-M50X

TABLE 6.1 – Liste des téléphones, des versions d'Android et des casques utilisés par les sujets. Certains sujets n'ont donné que la marque du casque, sans le modèle. Le sujet 20 n'a pas donné sa marque de casque.

l'expérience.

Enfin, l'expérimentateur recommande au sujet d'accomplir ses sessions tant que faire se peut au moment où il reçoit la notification par SMS. Cependant, pour anticiper les moments d'indisponibilité, il est aussi indiqué qu'il est possible de différer sa session dans la demi-journée, voire même de l'anticiper si besoin (faire sa session avant d'avoir reçu le SMS). Dans ce cas, dans la mesure du possible, l'expérimentateur demande au sujet de ne pas installer une routine de report à une heure qui serait toujours la même et dans le même contexte.

6.3.4 Débriefing

À l'issue de l'expérience, les sujets sont invités de nouveau à se rendre au laboratoire pour faire un bilan de l'expérience et recevoir leur défraiement. Le bilan se fait en deux étapes, d'abord en répondant à un questionnaire par écrit (disponible en annexe B.2), et ensuite en discutant oralement sur la base du questionnaire. Pour les échelles de notation du questionnaire écrit, les sujets ont pour instruction de les annoter avec un simple trait, sans mettre de chiffre. Enfin, les sujets reçoivent le cas échéant leur défraiement.

6.3.5 Hypothèses

Compte tenu de la difficulté d'émettre un jugement prédictif sur l'influence du contexte, nous formulons les hypothèses suivantes, relatives à l'influence du son binaural sur l'immersion, la mémorisation et la performance. Dans le cas des sessions binaurales donc, par rapport aux sessions mono, nous supposons obtenir :

— pour l'immersion, une réponse plus favorable aux quatre questions posées (meilleur sentiment de présence, son 3D plus souvent entendu, son qui contribue davantage au sentiment d'immersion et un contexte ressenti comme moins gênant) ;

— pour la mémorisation, un ordre de restitution des objets de mémoire plus fidèle à l'ordre des objets croisés et un son jugé plus utile pour ordonner ;

— pour la performance, un score global (distance parcourue et bonus collectés) supérieur.

6.4 Résultats

6.4.1 Introduction

Les résultats présentés dans cette section correspondent aux 70 sessions de tests effectuées par les 30 participants. Au total donc nous récoltons les informations de 2100 sessions. Pour des raisons techniques (problèmes de communication avec le serveur), les 12 premières sessions du sujet 2 sont perdues, réduisant les données à 2088 points.

Si les données contextuelles sont récoltées à chaque fois, les questionnaires d'immersion et de mémorisation n'apparaissent pas systématiquement (voir section 6.3.1). Nous avons pour chacun d'eux 896 et 894 entrées, et 298 entrées sans questionnaire.

Une session dure en moyenne 3min2s, dont 2min21s de jeu, 18s de réponse au questionnaire de contexte, et 23s de réponse au questionnaire de fin (immersion, mémorisation ou rien). En 70 sessions, les sujets auront donc passé en moyenne 3h33min sur l'application.

6.4.2 Observation du contexte

Les données contextuelles sont représentées sur des diagrammes circulaires en Figure 6.7. On observe d'emblée la prépondérance de certaines composantes contextuelles par rapport à d'autres. Pour le lieu, « À la maison » recueille 52% des réponses, tandis que « Au travail/À l'école » en recueille 27%. Pour le contexte social, « Seul(e) » et « Entouré(e) d'une ou plusieurs personnes connues » rassemblent respectivement 49% et 45% des réponses. Le contexte de mobilité voit « Assis(e)/Allongé(e) » réunir 78% des réponses. Pour le niveau d'occupation, les réponses sont plus équilibrées, on constate tout de même 26% et 21% pour les niveaux 2 et 3. Par ailleurs, 634 sessions (i.e., 30% des cas) se sont déroulées en étant seul, à la maison et assis ou allongé (avec des niveaux d'occupation divers). On constate donc que nombre de ces contextes ne sont pas spécifiques aux terminaux mobiles, en particulier pour le niveau de mobilité. Les résultats de localisation concordent par ailleurs avec ceux de [N. Liu, Y. Liu et X. Wang, 2010], i.e., 68,56% d'utilisation à la maison/dans un lieu pour dormir, 22,35% au travail/à l'école/dans un lieu de restauration et 9,09% dehors/dans le bus/sur la route.

6.4.3 Réponses au questionnaire d'immersion

Pour des raisons de lisibilité, les résultats des échelles à 4 degrés sont ramenés à une note entre 0 et 3, 0 correspondant à « Pas du tout » et 3 à « Tout à fait ». Les réponses aux questions relatives à l'immersion sont visibles sur la Figure 6.8. Aux première, troisième et quatrième questions les réponses vont dans le même sens : un meilleur sentiment de présence, une contribution du son plus importante à l'immersion et une gêne moins ressentie du contexte extérieur dans le cas des sessions binaurales. Les réponses à la deuxième question révèlent que les sujets ressentent du son 3D dans la majorité des cas, que le son soit rendu en binaural ou non.

Des ANOVA sont réalisées successivement sur les réponses aux questions 1, 3 et 4 avec le type de rendu sonore comme facteur intra-classe. Les résultats indiquent un effet significatif du type de rendu sur le sentiment de présence ($F(1, 29.03)=6.31$, $p<0.05$) et sur la contribution du son à l'immersion ($F(1, 29.04)=5.05$, $p<0.05$), mais pas sur la gêne ressentie du contexte extérieur ($F(1, 29.06)=1.09$, $p=0.30$).

Les notes obtenues sur la gêne ont un intervalle de confiance plus large que les autres. Il est intéressant d'observer sur la Figure 6.9 que les valeurs varient sensiblement avec

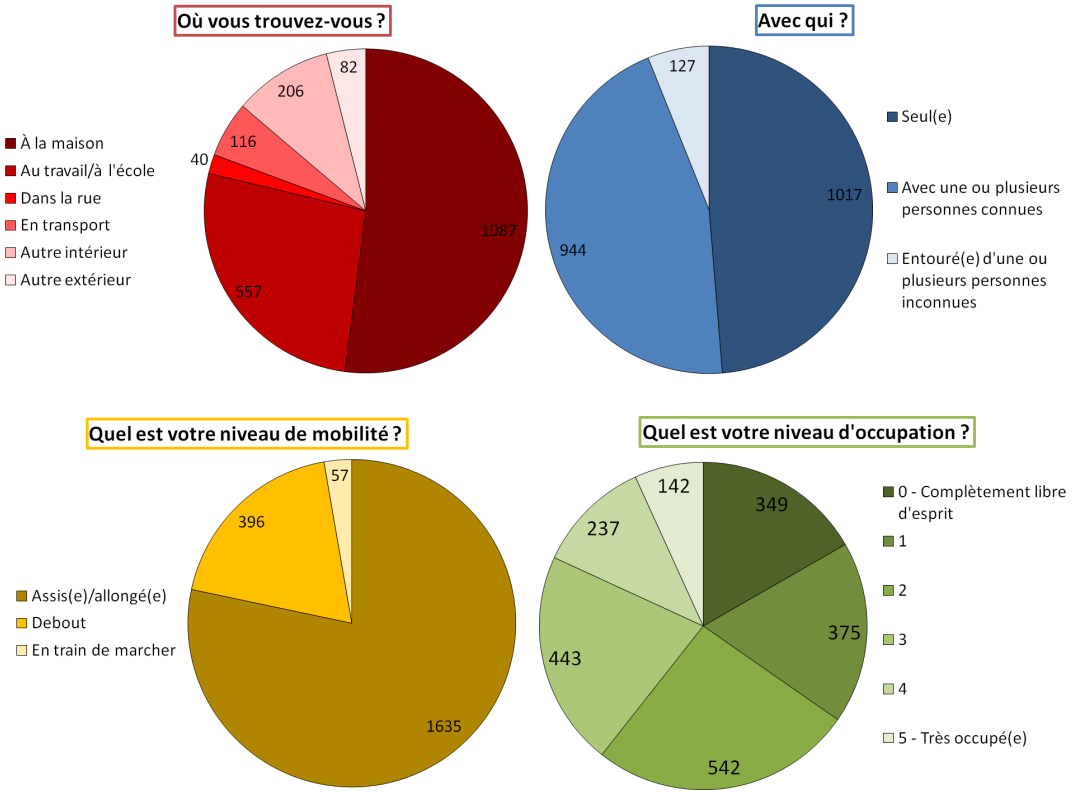


FIGURE 6.7 – Nombre d'occurrences des réponses aux quatre questions contextuelles, relatives au lieu (en haut à gauche) à l'entourage (en haut à droite), au niveau de mobilité (en bas à gauche%.) et au niveau d'occupation (en bas à droite).

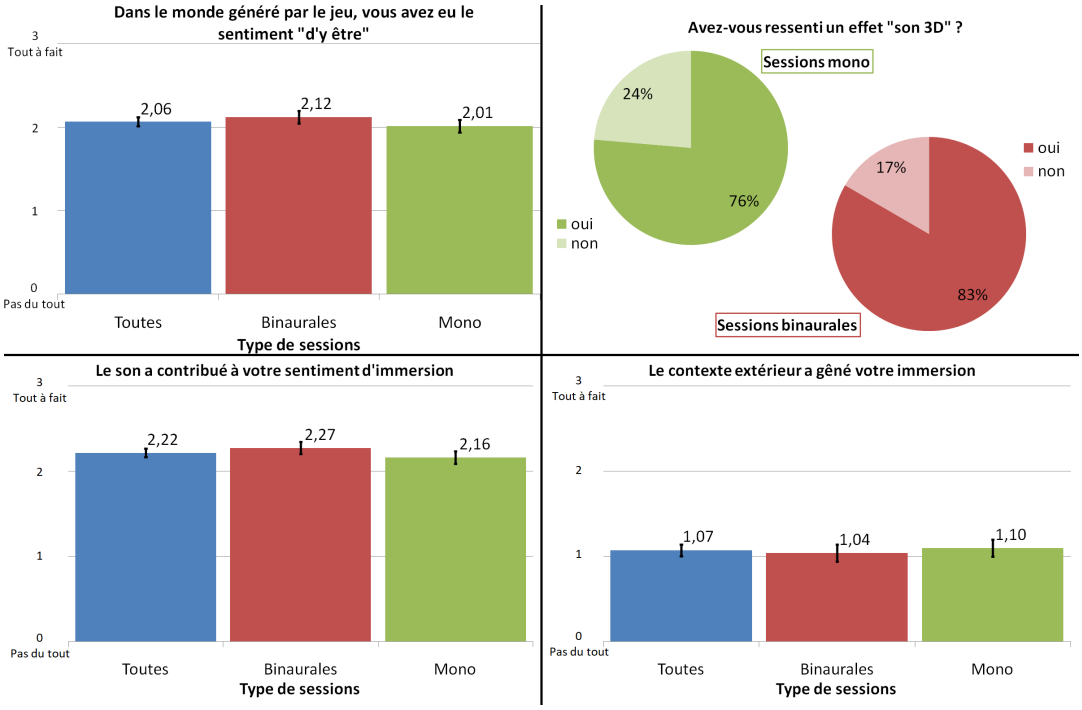


FIGURE 6.8 – Réponses moyennes obtenues aux questions sur l'immersion : a) le sentiment de présence (en haut à gauche), b) le ressenti du son 3D (en haut à droite), c) la contribution du son à l'immersion (en bas à gauche) et d) la gêne du contexte sur l'immersion (en bas à droite). Les barres noires verticales de a), b) et d) sont les intervalles de confiance à 95%.

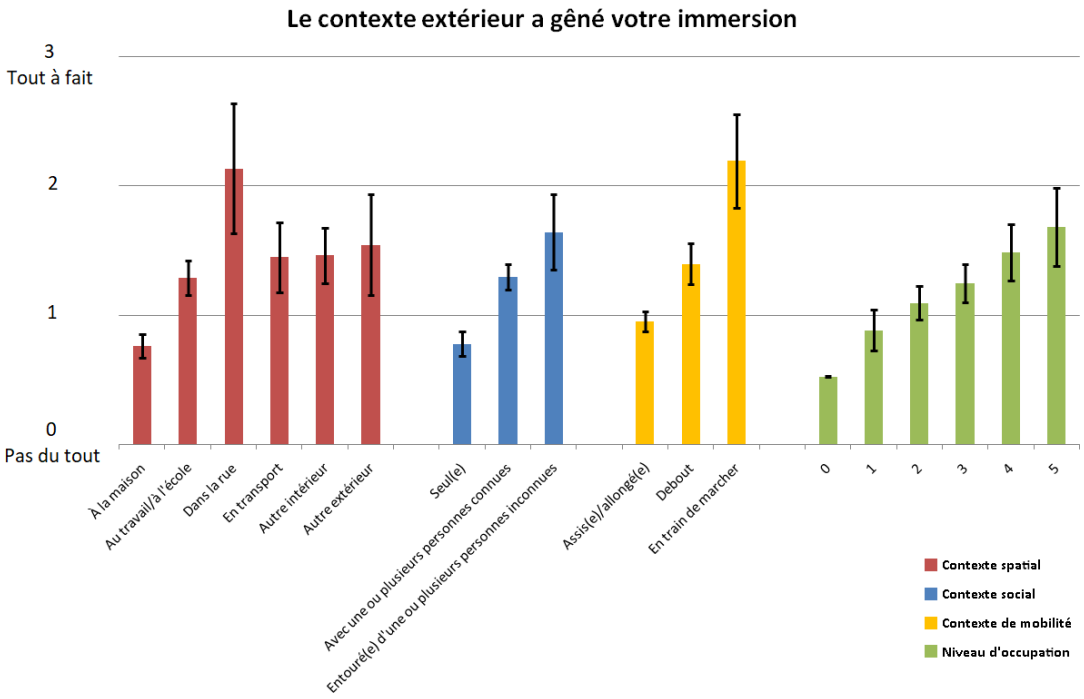


FIGURE 6.9 – Notes moyennes obtenues à l'assertion « Le contexte extérieur a gêné votre immersion » en fonction du contexte. Les barres verticales sont les intervalles de confiance à 95%.

le contexte. Le « bruit » introduit par ces facteurs d'influence sur les données peut être une des raisons expliquant l'absence de significativité du rendu sonore, malgré une amélioration des notes du même ordre de grandeur que pour les autres questions dans le cas des sessions binaurales. Par ailleurs, on remarque que les variations sont cohérentes avec le contexte : faible gêne ressentie lorsque le sujet est à la maison, seul, assis ou allongé, et l'esprit faiblement occupé ; à l'inverse, forte gêne lorsqu'il est dans la rue, entouré de personnes inconnues, en train de marcher, et l'esprit fortement occupé. Gardons à l'esprit que ces observations ne sont faites qu'à titre indicatif, dans la mesure où le contexte n'est pas un paramètre contrôlé et uniformément distribué au travers des sessions.

Enfin, un test du χ^2 est mené sur les réponses à la question « Avez-vous ressenti un effet "son 3D" ? » (réponses considérées comme variable dépendante, le type de rendu sonore étant la variable indépendante). Les résultats révèlent une forte présomption contre l'hypothèse nulle, avec une valeur de χ^2=6.81 et $p<0.01$. Il y a donc un effet statistiquement significatif du type de rendu sonore sur la réponse des sujets, i.e., les sujets semblent percevoir sensiblement plus du son 3D dans les sessions en binaural.

6.4.4 Résultats de la tâche de mémorisation

Calcul de distance ou de similarité entre deux séquences d'objets

Pour la tâche de mémorisation, le sujet fournit la liste des 7 derniers objets de mémoire, dans l'ordre où il se rappelle les avoir croisés. Notre objectif est de la comparer à celle des objets croisés dans le bon ordre. Il existe de nombreuses méthodes pour calculer la distance entre deux séquences. La plupart d'entre elles sont utilisées pour mesurer la similarité entre des chaînes de caractères (pour les auto-corrections de SMS sur smartphone par exemple). Nous identifions cinq calculs de distance ou de similarité qui peuvent être pertinents.

Le premier est la distance de Hamming : elle comptabilise simplement les éléments qui ne sont pas positionnés au même endroit entre les deux séquences. Dans notre cas, avec des séquences de 7 objets, elle renvoie une valeur comprise entre 0 (les deux séquences identiques) et 7 (aucun objet à la même place). C'est un calcul intuitif, mais il ne tient pas compte de l'ordre des objets ni de la distance qui les sépare. Par exemple, les séquences « trompette - machine à laver - piano - réveil » et « réveil - trompette - machine à laver - piano » donneraient une distance maximum, car aucun objet n'est à la même place. Pourtant, la trompette, la machine à laver et le piano sont restitués dans le bon ordre et à une faible distance de leurs positions d'origine. Étant donné que notre tâche de mémorisation est une tâche d'ordonnancement, il nous semble important de proposer des calculs qui tiennent aussi compte de ces critères.

La longueur de la plus grande sous-séquence commune permet de tenir compte de l'ordre. À l'inverse de la distance de Hamming elle mesure une similarité, et renvoie dans notre cas une valeur comprise entre 1 et 7. Elle néglige cependant la bonne ou mauvaise position de cette sous-séquence.

La distance de Levenshtein comptabilise le nombres d'opérations (substitution, insertion, suppression) nécessaires pour transformer une séquence en une autre. Par exemple, pour passer de « trompette - machine à laver - piano - réveil » à « réveil - trompette - machine à laver - piano », la distance sera de 2, car 2 opérations sont nécessaires :

— la suppression du réveil, pour passer de « trompette - machine à laver - piano - réveil » à « trompette - machine à laver - piano » ;

— puis l'insertion du réveil à la bonne place, pour passer de « trompette - machine à laver - piano » à « réveil - trompette - machine à laver - piano ».

Comme pour la plus petite sous-séquence commune, ce calcul favorise les séquences déjà proches dans leur ordre, mais néglige la proximité spatiale d'un élément par rapport à sa position d'origine. De même que pour Hamming, cette valeur est comprise entre 0 et 7 dans notre cas.

La distance de Damerau-Levenshtein intègre ce critère. Il s'agit du même calcul que Levenshtein, avec une opération supplémentaire autorisée : l'échange de deux éléments adjacents (la transposition). Par exemple, considérons la séquence originale « trompette - machine à laver - piano - réveil » et la séquence restituée par le sujet « trompette - piano - machine à laver - réveil ». Tandis que la distance de Levenshtein est de 2

(suppression de la machine à laver, puis insertion à la bonne place), la distance de Damerau-Levenshtein réduit cette distance à 1, car elle autorise à échanger le piano et la machine à laver. Si en revanche on considère une séquence restituée « trompette - piano - réveil - machine à laver », où la machine à laver a été décalée encore d'un cran vers la fin, bien que la distance de Levenshtein ne change pas, celle de Dameareau-Levenshtein augmente à 2, car une transposition simple ne permet plus d'obtenir la séquence d'origine. On favorise donc les éléments proches, mais seulement lorsqu'ils sont côte à côte.

La similarité de Jaro est une valeur comprise entre 0 (séquences très différentes) et 1 (séquences identiques). Le calcul combine le nombres de correspondances m entre les deux séquences et le nombre de transpositions t. Une correspondance est établie entre deux éléments identiques des deux séquences si leur éloignement est inférieur ou égal à :

$$\left\lfloor \frac{max(\mid s_1 \mid, \mid s_2 \mid)}{2} \right\rfloor - 1$$

où $\mid s_i \mid$ est la longueur de la séquence i.

On compare ensuite deux à deux les i-èmes éléments correspondants de s_1 et de s_2. Le nombre de transpositions correspond aux total de différences comptabilisées, divisé par 2.

Finalement, la distance de Jaro d_j entre deux séquences s_1 et s_2 est définie par :

$$d_j = \frac{1}{3}\left(\frac{m}{\mid s_1 \mid} + \frac{m}{\mid s_2 \mid} + \frac{m-t}{\mid s_1 \mid}\right)$$

Comparaison des distances entre sessions binaurales et sessions mono

La Figure 6.10 présente les distances moyennes obtenues pour toutes les sessions avec questionnaire de mémorisation, puis pour celles rendues en binaurales et celles rendues en mono. On observe que pour chaque type de distance, les valeurs sont toutes très proches et relativement hautes. Seules les distances représentées par la plus longue sous-séquence commune et Jaro présentent des valeurs relativement basses. Gardons cependant à l'esprit que la plus longue sous-séquence commune est un critère moins restrictif que les autres, puisqu'il favorise uniquement les suites d'objets identiques, indépendamment de leurs positions dans la liste. Par ailleurs, la distance de Jaro avantage fortement le fait que les deux listes possèdent les mêmes objets (impliquant un nombre de correspondances plus élevé). On remarque cependant que toutes les sessions binaurales ont une distance plus grandes que les sessions mono (ou à l'inverse une similarité plus faible), ce qui indiquerait une mémorisation moins bonne des objets dans le cas du binaural. Des ANOVA sont menées, avec chacune pour variable dépendante une des distances, et le type de rendu sonore comme variable intra-classe. Les résultats ne révèlent pas d'effet significatif du rendu sonore, quelle que soit la distance : Hamming ($F(1, 29.24)=0.10$, $p=0.76$), Levenshtein ($F(1, 29.22)=0.71$, $p=0.41$), Damerau-Levenhstein ($F(1, 29.21)=0.45$, $p=0.51$), plus longue sous-séquence commune ($F(1, 29.28)=2.89$, $p=0.10$) ou similarité de Jaro ($F(1, 29.21)=0.34$, $p=0.56$).

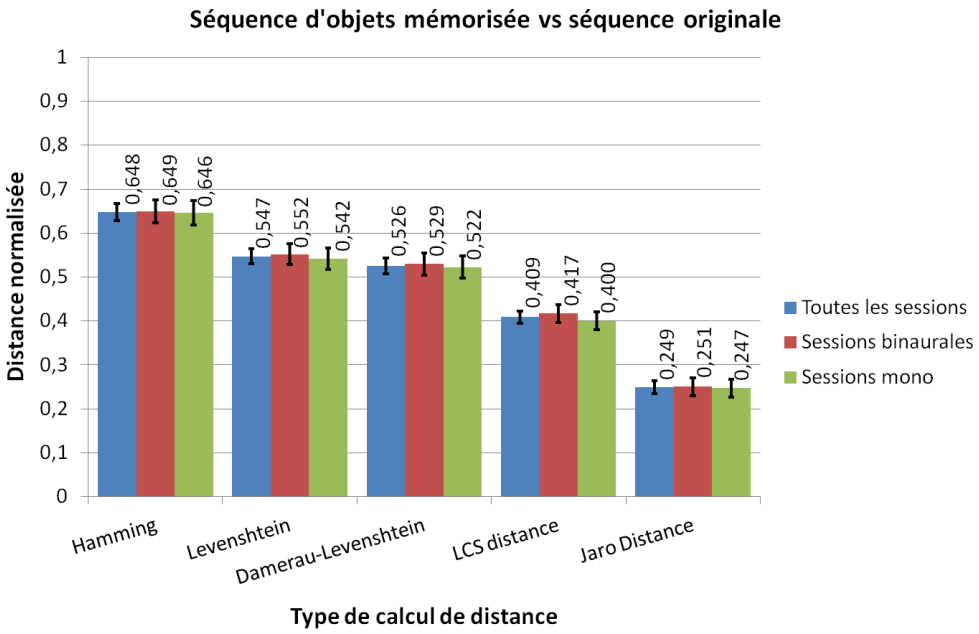


FIGURE 6.10 – Distances moyennes entre la séquence d'objets croisés par les sujets pendant la phase de jeu et la séquence restituée pendant le questionnaire de mémorisation, selon le type de rendu sonore. Pour faciliter la visualisation, les valeurs sont normalisées. La plus longue sous séquence commune et la similarité de Jaro ont été inversées, de façon à n'avoir que des représentations de distance. De ce fait, pour l'ensemble des critères observés, plus la valeur est basse, plus la mémorisation est bonne. Les barres verticales sont les intervalles de confiance à 95%.

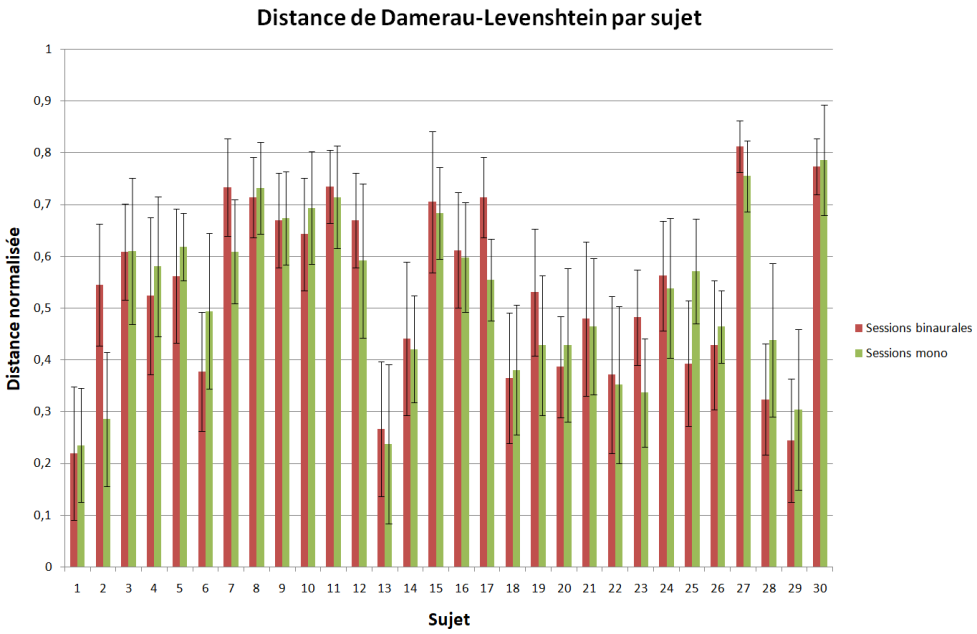


FIGURE 6.11 – La distance de Damerau-Levenshtein calculée sur les séquences de mémorisation, par sujet. Les barres verticales sont les intervalles de confiance à 95%.

Contrairement à la gêne du contexte extérieur ressentie par le sujet, qui montre des différences notables d'un contexte à l'autre, la mémorisation ne varie pas beaucoup en fonction du lieu, de l'entourage, de la mobilité ou de l'occupation du sujet. En revanche, nous remarquons une disparité entre les sujets. La Figure 6.11 montre la distance de Damerau-Levenshtein. Nous choisissons cet indicateur car il nous semble plus complet, mais des observations similaires ont été constatées avec les autres distances. On observe qu'en moyenne la moitié des sujets mémorisent mieux les séquences lors des sessions binaurales par rapport aux sessions mono. Cependant, l'importance des intervalles de confiance ne nous permet pas d'aller au delà de cette simple observation.

Contribution de l'auditif et du visuel à la mémorisation

La Figure 6.12 permet d'observer, selon le type de rendu sonore, les réponses moyennes aux deux assertions (« Le positionnement sonore des objets vous a aidé à classer les objets » et « Le positionnement visuel des objets vous a aidé à classer les objets ») sur l'aide des positions sonores ou visuelle des objets pour mémoriser. On remarque dans un premier temps que le son aide beaucoup plus les sujets à mémoriser que le visuel.

Par ailleurs, pour la modalité sonore, les réponses n'augmentent que très légèrement dans les sessions en binaural et augmentent plus visiblement pour la modalité visuelle. Cependant, une ANOVA menée sur les données, considérant successivement les réponses aux deux questions, avec le type de rendu sonore comme variable intra-classe, ne révèle aucun effet significatif du type de rendu sonore sur les réponses, modalité visuelle ($F(1, 29.19)=2.9$, $p=0.09$ comme sonore ($F(1, 29.18)=0.39$, $p=0.54$).

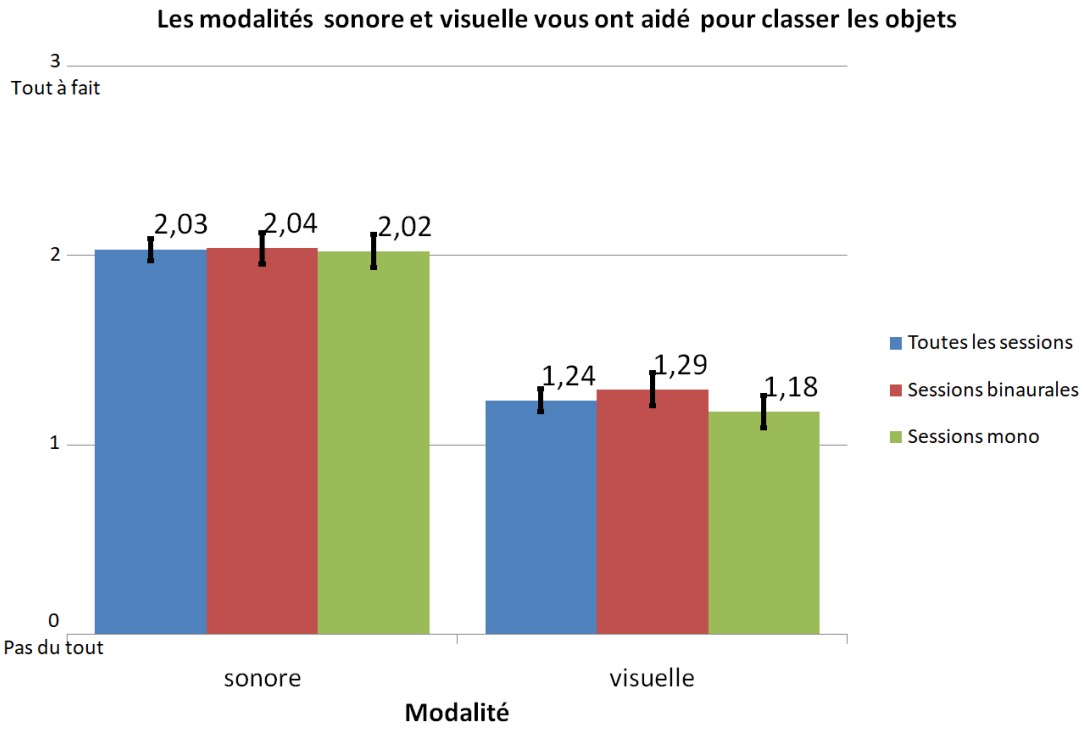


FIGURE 6.12 – Réponses moyennes données par les sujets aux assertions « Le positionnement sonore de ces objets vous a aidé à classer les objets » et « Le positionnement visuel de ces objets vous a aidé à classer les objets », ici présentées selon le type de rendu sonore de la session. Les barres verticales sont les intervalles de confiance à 95%.

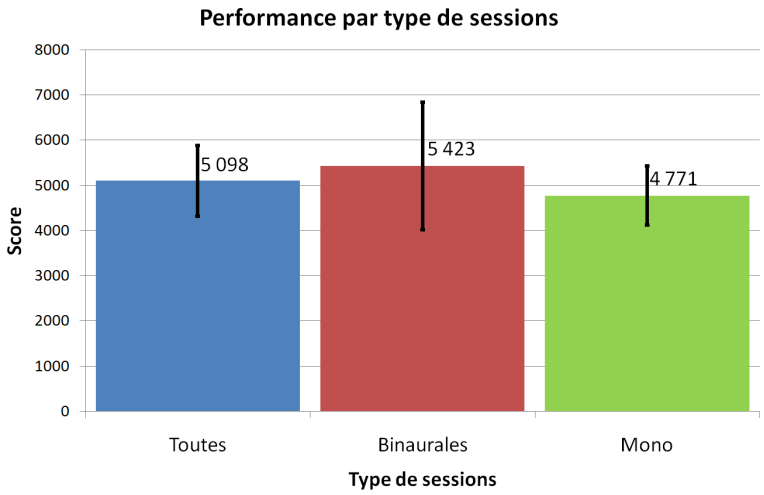


FIGURE 6.13 – Score moyen de performance obtenu selon le type de rendu sonore des sessions. Les barres verticales sont les intervalles de confiance à 95%.

6.4.5 Observation du score

Les scores moyens obtenus sont représentés selon le type de rendu sonore des sessions dans la Figure 6.13. On observe ici un score moyen meilleur pour les sessions en binaurales. Par ailleurs, la Figure 6.14 permet d'observer les scores obtenus par sujet selon le type de rendu sonore. On s'aperçoit que le sujet 8 (et dans une moindre mesure le sujet 19) possède un score moyen bien au dessus de tous les autres sur les sessions en binaural. Celui-ci détient en effet les trois scores les plus élevés de l'expérience (le premier dépassant les 660000 points, soit un facteur 100 par rapport à la moyenne), obtenus sur des sessions en binaural. Indépendamment du sujet 8, l'ensemble des autres sujets cumulent à eux tous une meilleure moyenne en mono qu'en binaural. Une nouvelle fois, une ANOVA sur les données, avec le type de rendu sonore en variable intra-classe, suggère l'absence d'effet significatif du rendu sonore sur le score ($F(1, 28.05)=0.18$, $p=0.68$).

La Figure 6.15 représente les scores par contexte. Comme pour l'immersion, les scores varient sensiblement et sont plus élevés dans des situations propices à la concentration : à la maison, assis ou allongé, seul, à un niveau d'occupation bas. Les scores élevés en transport et au niveau d'occupation 4 s'expliquent en partie par le fait que les plus hauts scores détenus par les sujets 8 et 19 ont été faits dans ces contextes.

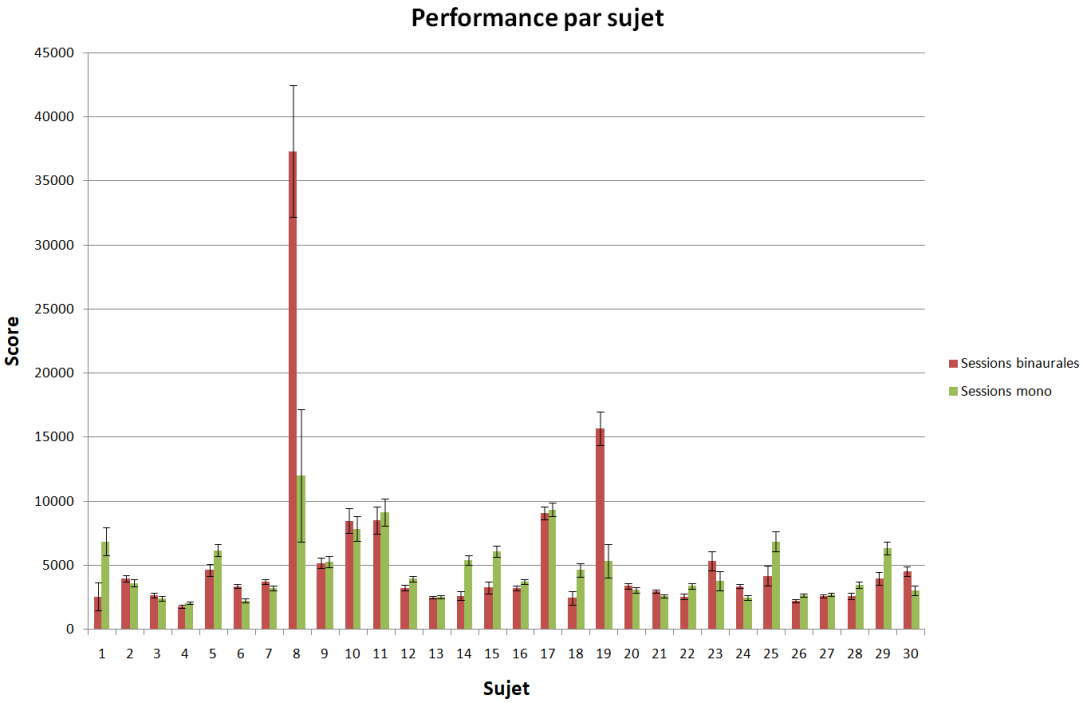


FIGURE 6.14 – Score moyen de performance obtenu pour chaque sujet selon le type de rendu sonore des sessions. Les barres verticales sont les intervalles de confiance à 95%.

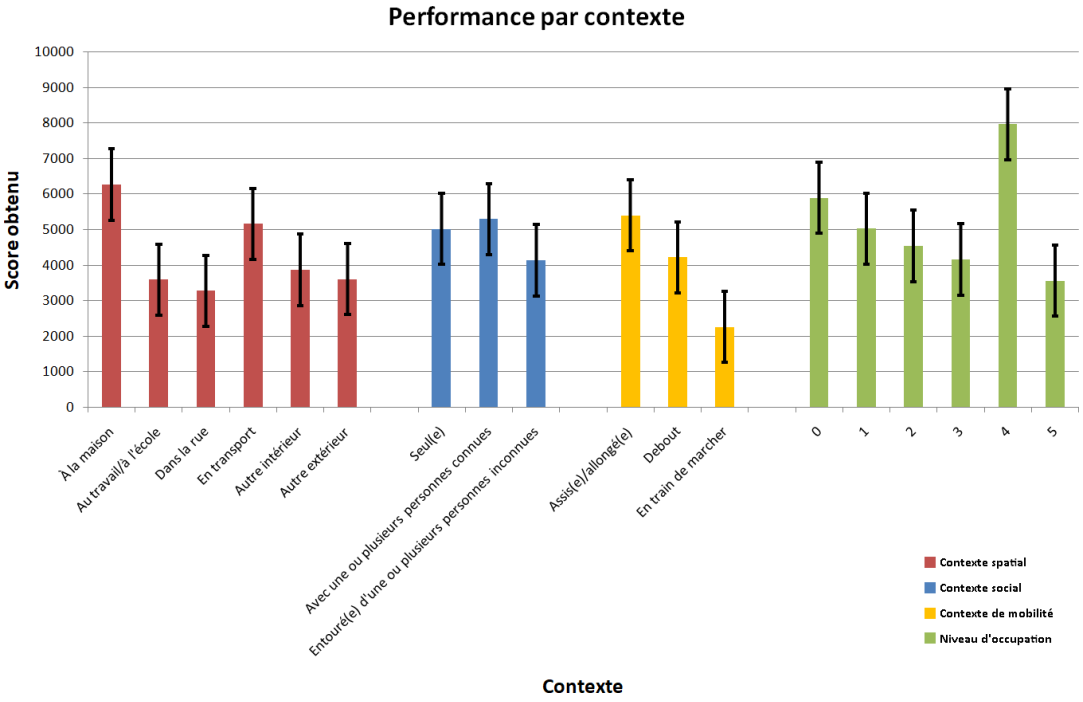


FIGURE 6.15 – Score moyen de performance obtenu selon le type de contexte. Les barres verticales sont les intervalles de confiance à 95%.

6.5 Discussion

6.5.1 Sur les résultats

Les résultats sur l'immersion, la mémorisation et le score sont disparates. Sur les quatre questions du questionnaire de l'immersion, trois ont vu leurs réponses significativement influencées par le binaural. Il semble que le son binaural renforce le sentiment de présence du joueur et qu'il contribue davantage à son immersion que le son mono. Par ailleurs, le son 3D est ressenti plus souvent lors des sessions binaurales, indiquant que les sujets perçoivent consciemment la spatialisation sonore. En revanche, les sujets ne perçoivent pas le contexte extérieur comme moins gênant grâce au binaural. Nous avons montré que les différents contextes représentés modifiaient cette sensation de gêne, et que les variations induites pouvaient diluer l'effet du binaural. Gardons cependant à l'esprit que le pouvoir isolant du binaural sur le contexte extérieur est sans doute dû en premier lieu à l'utilisation du casque, qui était aussi utilisé dans les sessions mono de notre expérience.

En ce qui concerne la mémorisation, les réponses des sujets suggèrent que le son est plus utile que le visuel pour mémoriser. Cela s'explique sans doute par le fait que dans un *Infinite Runner*, le décor défilant vite, le temps d'apparition des objets à l'écran est court. Dans ces conditions, une fois l'objet passé hors-champ, le fait d'entendre encore un moment le son passé derrière soi devient probablement un atout pour la mémorisation. Cependant, le binaural ne renforce significativement pas cette aide. De même concernant la tâche de mémorisation elle-même, le rendu sonore n'influence pas significativement les résultats. Les différentes valeurs de distance calculées nous laissent à penser que la tâche de mémorisation était particulièrement difficile, observation confirmée par les témoignages oraux de plusieurs sujets lors du bilan : retenir l'ordre des 7 derniers objets croisés avait été une tâche difficile en soi, rendue d'autant plus ardue par le fait que ces 7 objets venaient parfois compléter une liste totale beaucoup plus longue. Pour cette raison, certains sujets ont même annoncé avoir régulièrement fait exprès de perdre une fois les 7 premiers objets croisés, pour leur éviter de compliquer davantage la tâche.

Au delà de cette question de difficulté, il pourrait être intéressant dans une étude ultérieure de mettre en relation les résultats de mémorisation avec l'attention portée à la tâche : dans le débriefing, certains sujets ont reporté oralement avoir focalisé leur attention sur la tâche de mémorisation plutôt que sur le but du jeu (montrant au passage que notre effort à gommer l'effet d'anticipation des sujets en alternant les questionnaires n'a pas toujours fonctionné). L'attention des sujets a certainement joué un rôle majeur dans la réussite de cette tâche, et a peut-être un effet croisé avec le type de rendu sonore.

Pour finir sur la mémorisation, notons que les objets étaient répartis sur la route, seulement à droite ou à gauche du sujet. On sait que le binaural non-individualisé favorise surtout la spatialisation latérale, raison qui nous a poussés à limiter les positions, pour mettre en avant sa spécificité et son efficacité de spatialisation. Cependant cette volonté rentre quelque peu en conflit avec le principe de l'art de la mémoire (voir Section 5.4.2), qui réclame de se représenter l'objet mentalement à une position

unique. Un sujet nous a même explicitement dit avoir utilisé l'art de la mémoire pour cette tâche, mais en plaçant les objets à des positions complètement indépendantes de celles rencontrées dans le jeu. Placer les objets en des endroits plus variés, tout autour de l'avatar, aurait peut-être permis une meilleure association entre objet et position.

Enfin, concernant les scores, aucun effet significatif du rendu sonore n'a été révélé statistiquement. Les facteurs d'influence semblent trop nombreux dans notre expérience : contextes variables et distribués aléatoirement, ne nous permettant pas d'approfondir l'étude des scores au cas par cas, en comparant par exemple les sessions binaurales dans la rue avec les sessions mono dans la rue. En plus de ça, plusieurs parties ont été interrompues volontairement, rendant le score non représentatif de l'aptitude du joueur : jeu parfois jugé trop facile par certains sujets, qui mettaient fin à leur partie pour éviter d'y passer trop de temps ; ou partie limitée dans le temps à cause des conditions extérieures (trajet dans un transport arrivant à son terme, pause au travail limitée, etc.) ; ou encore interruption volontaire une fois les 7 objets de mémorisation croisés. Dans ces conditions, il est difficile de conclure sur nos résultats.

6.5.2 Sur la méthodologie, les problèmes de mise en œuvre

La mise en œuvre concerne le déroulement concret de l'expérience. Dans un travail qui expose une hypothèse scientifique à vérifier, elle ne fait pas toujours, voire rarement, l'objet d'une discussion. Pourtant, la recherche étant limitée dans le temps et les moyens, les limitations pratiques sont aussi une façon de justifier le choix de la méthode. Plus encore, dans notre cas, l'expérience a été menée avec l'intention de conférer aux résultats une validité externe la plus grande possible. Pour cette raison, la méthode ESM a été choisie, et les sessions des sujets ont été soumises à de nombreux facteurs d'influence non récoltés dans cette expérience, potentiellement impactant sur la validité interne des données. Dans la section qui suit, nous esquissons les principaux aspects des difficultés techniques rencontrées, liées à l'ESM, et des quelques facteurs d'influence possibles résultant.

Sélection des sujets

Une difficulté rencontrée dans cette expérience concerne la sélection des sujets et l'utilisation de leur téléphone personnel. Notre application n'ayant pas été testée sur l'ensemble des marques et modèles de téléphone existants, des problèmes récurrents nous ont privé de plusieurs volontaires, sur le téléphone desquels l'application ne fonctionnait pas, n'envoyait pas correctement les données sur le serveur (problème d'identification), cessait de fonctionner dès le lancement, ou après quelques secondes de jeu. Au total, 10 sujets sont venus se faire installer l'application sans succès.

Motivation et interruptions des sujets

Par ailleurs, la disponibilité et la motivation des sujets sont aussi un problème de taille dans ce genre de protocole. Sur les 30 sujets qui sont allés au bout des 70 sessions,

seulement 2 ont terminé dans les 35 jours impartis. Le nombre moyen de jours de retard est de 14,7 jours. Étant donné la disparité des retards entre les sujets, il est difficile d'analyser les répercussions de ces rallonges de temps, entrecoupées d'interruptions, sur les résultats. Par ailleurs, ces retards ont entrainé des difficultés d'ordre logistique, nécessitant de prolonger manuellement et individuellement les envois de SMS. Un système de notification automatique tenant compte de cet aspect est donc plus que recommandé avant le déploiement d'une telle expérience.

Malgré l'envoi de SMS journaliers, 7 autres sujets ont abandonné l'expérience en cours de route. Le paiement échelonné est censé constituer un moteur dans ces cas là, mais il doit alors être suffisamment élevé et réparti tout au long de l'expérience, ce qui n'était sans doute pas assez le cas ici. Par ailleurs, 2 sujets supplémentaires ont dû interrompre leur expérience après avoir cassé leur téléphone, et 1 autre pour avoir modifié les paramètres de son téléphone à plusieurs reprises pendant l'expérience, effaçant les fichiers de sauvegarde de l'application.

6.5.3 Contrôle des sujets

Une autre difficulté concerne le contrôle des sujets. Il est en effet impossible de vérifier que les sujets font leurs sessions comme il leur a été demandé. Pendant le bilan, et malgré les instructions insistantes données à ce sujet (à l'écrit via l'application et à l'oral), 9 sujets ont avoué avoir accompli quelques sessions sans casque (avec le son du téléphone ou sans son), 9 autres (dont certains se recoupent avec les précédents) avec des écouteurs intra-auriculaires, et 1 sujet a déclaré n'avoir pas fait attention au sens du casque dans les premières sessions (sans mentionner les possibles inversions de casque non détectées). Il nous est évidemment impossible d'identifier les sessions concernées.

De même, nous ne pouvons nous assurer que le sujet a bien accompli lui-même toutes ses sessions (néanmoins aucun cas de ce genre ne nous a été reporté). La session libre était également prévue pour permettre de faire jouer d'autres personnes (à la demande de certains sujets), quoique très peu de sessions libres ont été effectuées (aucune la plupart du temps).

Quoi qu'il en soit, toutes ces données aberrantes constituent un bruit de fond inévitable dont nous ne pouvons qu'espérer qu'il n'affecte pas les effets substantielles. À l'avenir, des mécanismes doivent être envisagés pour limiter au mieux ce genre d'entorse au protocole (par exemple, pour empêcher les problèmes de casque, un verrou auditif avant chaque session, qui se débloque après une tâche de localisation auditive simple, où le sujet indique dans quelle direction, gauche, droite ou centrale, il entend une succession de bips sonores).

6.6 Conclusion

Dans cette expérience, nous avons mesuré l'apport du binaural à une expérience de jeu vidéo mobile de type *Infinite Runner*. Des sessions de jeu avec binaural ont été comparées à des sessions de jeu en mono selon trois critères : le sentiment d'immersion, la mémorisation d'objets et la performance de jeu. Les résultats ont révélé une amélioration significative de l'immersion grâce au binaural, mais un apport plus indécis concernant la mémorisation et la performance. À tout le moins, aucune contre-indication n'a été trouvée concernant l'utilisation du binaural. Cette expérience suggère donc un apport positif du binaural à la qualité d'expérience, justifiant son emploi dans un jeu vidéo de ce type.

Plusieurs perspectives sont envisageables, du point de vue de l'apport du binaural d'abord. La comparaison avec un rendu monophonique a sans doute permis de mettre en avant les spécificités du son spatialisé, mais peut-être au détriment des propriétés particulières du binaural. Par exemple, les avantages liés à la latéralisation sont peut-être tout autant valables avec de la stéréophonie. Il serait par conséquent intéressant de prolonger l'expérience en ajoutant une comparaison avec d'autres rendus sonores, en particulier la stéréophonie.

Une autre perspective, plus générale, serait bien sûr de mener l'expérience avec d'autres applications, voire même de décliner la même application sur d'autres supports (e.g., ordinateur ou télévision). Cela permettrait de mesurer encore davantage la pertinence des trois critères, immersion, mémorisation et performance et éventuellement de les adapter selon le type d'application. Par exemple, la performance, ici difficile à exploiter, pourrait se révéler autrement plus parlante dans le cas d'un jeu où la localisation d'objets fait partie intégrante des mécaniques de jeu.

Enfin, du point de vue méthodologique, la méthode ESM nous a permis de mener une expérience hors les murs et de collecter des données proches d'une utilisation réelle de smartphone, bien que soumises à de nombreux facteurs d'influence non-contrôlés. Les expériences à suivre, fondées sur celle-ci, devront utiliser ce défrichage expérimental en renforçant les points positifs (e.g., la diversité des contextes, la possibilité de collecter plus de données) et en atténuer au mieux les effets indésirables (e.g., réduire les difficultés logistiques, en préparant plus en amont le déploiement, ou réduire la distribution erratique des contextes, en intégrant par exemple des modules *context-aware*).

Une suite intéressante à cette expérience serait également de la reproduire, *mutatis mutandis*, dans des conditions expérimentales de laboratoire, et de comparer les résultats, de façon à confronter les deux approches, entre validité externe et validité interne. Nous signalons que cette expérience a déjà été menée, mais non incluse dans cette thèse par manque de temps. Elle fera cependant l'objet d'une publication indépendante.

Chapitre 7

Conclusion

7.1 Contributions

L'objectif initial de cette thèse était de mesurer l'apport du binaural dans les applications mobiles audiovisuelles. Au vu de l'ampleur de la tâche, notre approche a consisté à circonscrire petit à petit la problématique pour aboutir à une expérience réalisable dans le temps imparti et avec les moyens disponibles. Pour ce faire, un des fils conducteurs a été de considérer l'expérience binaurale sur mobile telle qu'elle pourrait être vécue dans une application audiovisuelle déployée auprès du grand public. En cela, ce travail s'inscrit dans la tendance actuelle visant à traiter l'expérience non plus comme un système isolé, déplacé dans un contexte idéal de laboratoire, mais comme réinsérée au cœur des facteurs d'influence susceptibles de l'altérer, et pourvue des attributs auxquels un utilisateur lambda aurait accès.

Nous avons considéré dans un premier temps l'objet audiovisuel seul, dépourvu de contexte applicatif et de contexte d'utilisation précis, pour nous permettre d'étudier la façon dont il est perçu par l'être humain. Nous avons proposé un tour d'horizon de la technologie binaurale, en envisageant en particulier le problème de l'individualisation, son inaccessibilité auprès du grand public dans l'état actuel des choses et l'éventualité de s'en passer dans un contexte audiovisuel. Nous avons ensuite évoqué l'association d'un son spatialisé avec un visuel, l'effet ventriloque qu'elle engendrait, en le considérant non pas comme un effet indésirable, mais au contraire comme une opportunité de déployer la scène auditive en dehors de l'écran, sans briser ses liens avec la scène visuelle. Tous ces éléments ont été exposés dans le chapitre 2.

L'état de l'art sur l'effet ventriloque ne traitant pas spécifiquement le cas du son binaural couplé à un visuel sur mobile, nous avons proposé une expérience dans le chapitre 3, dans laquelle on détermine la fenêtre d'intégration auditivo-visuelle horizontale en utilisant la méthode du PSSA. Les résultats ont montré que l'individualisation des HRTF n'avait pas d'effet significatif sur la fenêtre, suggérant la possibilité de s'en affranchir, dans le cas où son et image sont associés de façon diégétique et univoque. Un large décalage en élévation entre son et image n'altère pas non plus significativement sa taille, nous permettant d'envisager sereinement les variations imprévisibles de hauteur du téléphone lorsque l'utilisateur le tient face à lui dans ses mains. Enfin, un décalage systématique de la fenêtre par rapport à la position du visuel nous laisse penser que les sujets adoptent un point d'écoute déplacé sur la caméra de la scène

virtuelle, attitude que nous avons qualifiée de processus de projection.

Dans le chapitre 4, nous nous sommes penchés sur la qualité sonore, la qualité d'expérience et les facteurs d'influence du contexte dans l'état de l'art. Cette partie a été l'occasion de replacer l'objet audiovisuel au sein de l'expérience mobile, et de mettre en lumière toute la complexité du domaine de l'évaluation de qualité, allant de la sélection d'attributs propres à représenter correctement ce qu'on veut évaluer de l'objet aux méthodes d'évaluation en elles-mêmes. En particulier, nous avons confronté deux approches du déploiement expérimental, une orientée validité interne (contrôle des facteurs d'influence et des biais expérimentaux) et l'autre orientée validité externe (résultats pertinents au delà de l'expérience).

Les chapitres 5 et 6 présentent l'expérience visant à mesurer l'apport du binaural. Le premier redéfinit la problématique de la thèse à la lumière de l'état de l'art et oriente les choix expérimentaux. Ainsi, dans le second, nous avons restreint l'étude à une application unique, un jeu vidéo de type *Infinite Runner*, et nous avons mesuré l'apport du binaural selon trois attributs, l'immersion, la mémorisation et la performance du sujet. Nous avons proposé un protocole expérimental fondé sur la méthode ESM, permettant de mener notre expérience au plus proche des cas d'utilisation réels. Malgré le bruit engendré par les nombreux facteurs d'influence sur les données collectées, les résultats ont révélé un apport significatif du binaural en termes d'immersion par rapport à une version de référence monophonique de l'application, justifiant l'emploi de cette technologie dans des conditions plausibles d'utilisation par le grand public (absence d'individualisation des HRTF, casques audio et smartphones de gammes variées, contextes multiples, etc.)

7.2 Perspectives

Ce travail de thèse offre de nombreuses perspectives. Parmi les choix tranchés que nous avons faits, et qui appellent à une étude plus poussée, il y a celui de l'association entre les sources sonores et les sources visuelles de la scène. Dès le chapitre 2, nous nous sommes contentés d'envisager un son associé à une image visible à l'écran, les deux étant temporellement synchronisés et sémantiquement cohérents. Pourtant, de nombreux autres situations auraient pu être envisagées, sans doute même certaines qui auraient pu bénéficier du binaural d'une façon différente (nous évoquions par exemple dans le chapitre 5 la spatialisation d'un son non-diégétique, créant un nouvel espace virtuel coexistant avec l'espace audiovisuel de la scène) et entraîner une réflexion nouvelle sur la perception de l'espace et des sources qui le compose. Un approfondissement de l'étude de la scène audiovisuelle et de ses possibilités, telles qu'elles ont été exposées par exemple dans [Chion, 2013] ou [Farnell, 2010], gagnerait donc à être fait.

Une autre ouverture possible concerne la variation des différents paramètres de l'expérience ESM. Changement d'application, mesure de l'apport du binaural selon d'autres termes (avec la possibilité de mener une étude pour déterminer les attributs propres à qualifier l'expérience audiovisuelle sur mobile avec du binaural), comparaison avec d'autres systèmes de restitution sonore, ou avec d'autres supports que le mobile (la

même application sur ordinateur par exemple), une mesure d'autres facteurs d'influence, etc., toutes ces possibilités participeraient à construire progressivement la validité interne de ces méthodes expérimentales hors les murs, tout en maintenant leur validité externe.

Annexe A

Documents de l'expérience PSSA

A.1 Instructions d'accueil

Notice d'information
Objectif et description de la recherche

L'expérience à laquelle vous allez participer est destinée à l'étude de la perception simultanée de stimuli visuels et auditifs. Les résultats recueillis seront analysés statistiquement.

Vous arriverez dans une pièce dans laquelle on vous demandera de vous asseoir sur une chaise, en face d'un téléphone mobile dont la position est fixe. Nous vous demanderons ensuite de régler votre distance au téléphone, puis votre élévation de manière à ce que l'inclinaison de votre regard forme un angle de 39° avec l'horizontale. Un fil tendu à 39° entre le téléphone et le mur vous aidera à vous positionner.

L'expérimentateur vous remet alors un contrôleur bluetooth qui vous permettra d'agir avec le téléphone. L'expérimentateur vous indiquera le bouton sur lequel appuyer. En aucun cas vous ne devrez toucher l'écran du téléphone.

Description du test :

Le test comprend deux blocs de séquences composées d'un stimulus sonore spatialisé diffusé sur casque d'écoute simultanément à un stimulus visuel projeté sur l'écran du téléphone. Dans les deux blocs, le stimulus auditif se déplace horizontalement, soit du centre vers la périphérie (droite ou gauche), soit de la périphérie vers le centre, tandis que le stimulus visuel est placé de manière fixe sur le téléphone, soit à droite soit à gauche. Chaque trajectoire peut avoir lieu dans le plan horizontal du téléphone (-39° par rapport à votre tête), soit dans un plan horizontal au-dessus de votre tête, à 70°.

Dans un bloc, le stimulus visuel est un hélicoptère dont les pales sont en mouvement, et le stimulus auditif est un bruit de rotor.

Dans un autre bloc, le stimulus visuel est un halo lumineux, et le stimulus auditif est un train de bruits blancs.

Chaque séquence est précédée d'un bip sonore indicatif.

Consigne :

Pour chaque séquence, votre tâche consiste à **fixer des yeux le stimulus visuel (axe du rotor de l'hélicoptère, ou centre du halo)** et à **indiquer l'instant dès lequel vous percevez** celui-ci **aligné** (même direction)**, voire confondu** (formant un unique événement) **avec le stimulus auditif**.

Pour indiquer l'instant où vous percevez les 2 stimuli alignés, il vous suffit d'**appuyer aussitôt sur le bouton du contrôleur bluetooth**. Dès cet instant, le stimulus visuel disparaît et le stimulus auditif s'interrompt. La séquence suivante démarre automatiquement quelques secondes plus tard. Il vous est demandé de **toujours garder le contrôleur dans la même main**.

Déroulement du test :

Le test se déroule en deux blocs dont l'ordre vous sera spécifié au dernier moment par l'expérimentateur :

Bloc A : visuel d'hélicoptère fixe et bruit de rotor en déplacement (droite ou gauche)

Bloc B : visuel d'halo lumineux fixe et train de bruits blancs en déplacement (droite ou gauche)

Chacun des deux blocs du test est précédé d'une session d'apprentissage, constituée de 8 séquences (environ 2 minutes). Le bloc proprement dit comprend 64 séquences et dure environ 13 minutes. Une pause vous sera proposée tous les 20 stimuli environ.

Au total, le test dure donc 30 minutes, sans compter les pauses.

Nous vous remercions de votre participation.

A.2 Questionnaire de débriefing

Date : **Session :** Bruit blanc / Hélico en premier
Sujet id : **Age :** **Sexe :**
Droitier ou Gaucher ?
Lunettes / lentilles : **correction :**
Problèmes d'audition ?
Distance au téléphone :

1) **Quelle session vous a semblé la plus difficile, l'hélicoptère ou le bruit blanc ?**

Pourquoi?
.....
.....
.....

2) **Pour chacune des sessions, pouvez-vous évaluer la difficulté globale de la tâche sur l'échelle ci-dessous :**

Hélicoptère :

Facile **Difficile**

1 **2** **3** **4** **5**

Bruit blanc :

Facile **Difficile**

1 **2** **3** **4** **5**

3) **Sur une échelle de 1 à 5, à combien évaluez-vous la synchronisation entre votre perception du point d'alignement et votre geste de validation ?**

1 -> Geste très en avance sur la perception (forte anticipation)
2 -> Geste légèrement en avance sur la perception (légère anticipation)
3 -> Geste parfaitement synchronisée avec la perception
4 -> Geste légèrement en retard sur la perception
5 -> Geste très en retard sur la perception

4) **Avez-vous bien perçu le son à l'extérieur de votre tête et devant vous ?**
.....
.....

5) **La distance du son vous paraissait-elle en adéquation avec la distance du téléphone ?**

.....
.....

6) **Avez-vous perçu les deux niveaux d'élévation des stimuli sonores ?**

.....
.....

7) **Avez-vous ressenti une gêne à indiquer un point d'alignement lorsque les stimuli sonores étaient au-dessus de vous ?**

.....
.....

8) **Avez-vous d'autres remarques ?**

.....
.....
.....
.....
.....
.....
.....
.....

Merci pour votre participation !

Annexe B

Documents de l'expérience ESM

B.1 Instructions d'accueil

Nous présentons ci-dessous les instructions lues par le sujet avant de démarrer l'expérience. Après l'installation de l'application, le sujet repartait avec les instructions, lui permettant de les consulter à tout moment si nécessaire.

Notice d'information
Objectif et description de la recherche

Introduction

Dans cette expérience, vous allez jouer à un jeu vidéo mobile. L'expérience se déroule sur 5 semaines (35 jours), pendant lesquelles vous exécuterez au total 70 sessions de jeu, à raison de 2 par jour. Les résultats seront envoyés au fur et à mesure sur une base de données distante, puis analysés statistiquement.

Accueil au laboratoire

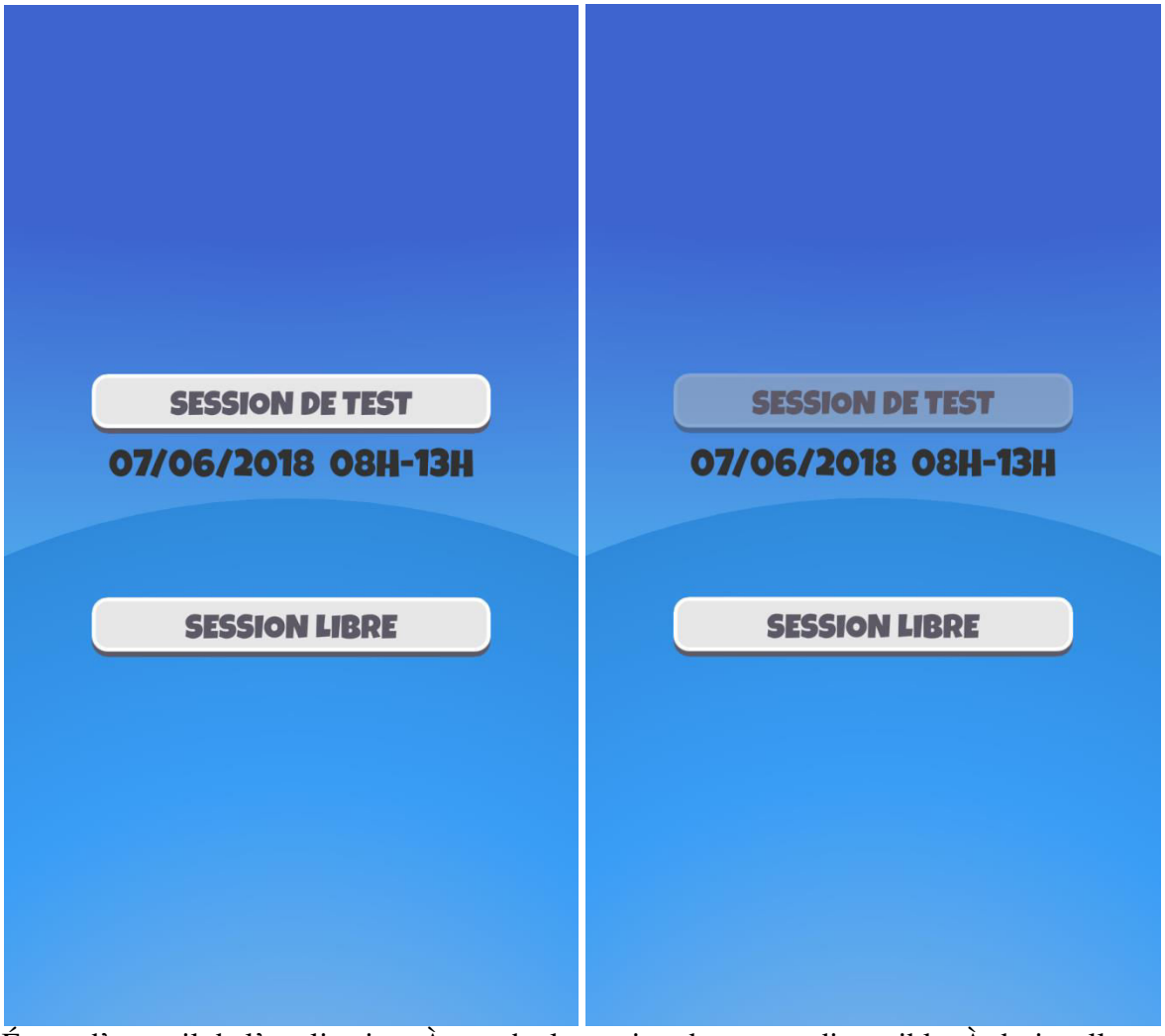
Aujourd'hui, l'expérimentateur va installer le jeu sur votre téléphone. Il collectera également des informations générales vous concernant (âge, habitudes de jeu, etc.) et répondra à vos éventuelles questions.

Organisation des sessions de jeu

Les sessions commenceront dès le lendemain. Chacune des deux sessions aura lieu à une heure aléatoire, une le matin entre 8h et 13h, et l'autre l'après-midi entre 14h et 18h. Vous recevrez à chaque fois un sms pour vous notifier du début de la session. Dans la mesure du possible, votre tâche sera de jouer dès sa réception. Si vous ne pouvez pas (à cause d'un empêchement quelconque), il vous est demandé alors de jouer dès que possible dans la tranche horaire correspondante (avant 13h pour une session le matin, avant 18h pour une session l'après-midi).

Si malgré tout vous ne parvenez pas à accomplir votre session dans la demi-journée impartie, celle-ci sera reportée à un jour ultérieur. Si vous avez manqué une session du matin, la session de l'après-midi est également reportée. Chaque session manquée rallonge donc d'une journée la durée totale de l'expérience.

Lorsque vous lancez l'application, vous arrivez sur l'écran ci-dessous. Deux types de sessions s'offrent à vous, la session de test et la session libre.



Écran d'accueil de l'application. À gauche la session de test est disponible. À droite elle est indisponible.

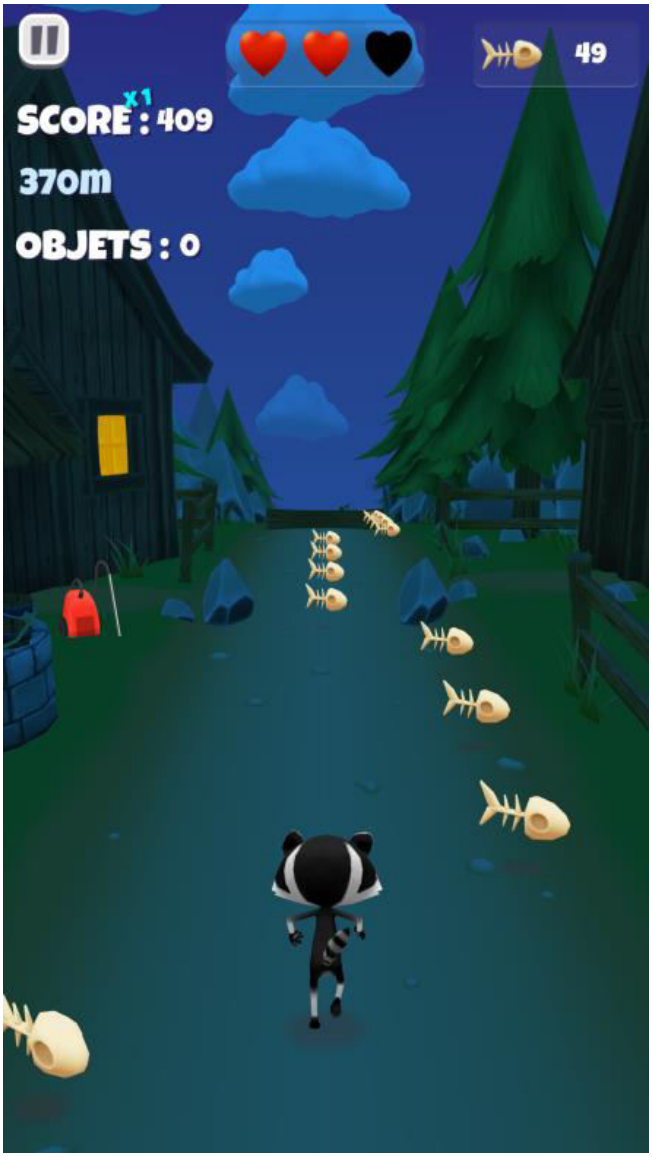
La session de test

C'est le type principal de session de l'expérience, celui que vous devez accomplir deux fois par jour. Une fois que la session de la demi-journée est exécutée, le bouton se grise et devient indisponible jusqu'à la prochaine session (à la date indiquée en dessous).

Une session se déroule en trois phases : un premier questionnaire sur votre contexte d'utilisation (durée moyenne estimée à 10 secondes), une phase de jeu (durée moyenne estimée à 4 minutes) et un second questionnaire sur différents aspects du jeu (objets rencontrés et immersion, durée moyenne estimée à 30 secondes).

Le premier questionnaire porte sur votre contexte d'utilisation. Quatre questions vous sont posées, toujours les mêmes, et concernent votre contexte au début de la session.

Vient ensuite la phase de jeu. Vous incarnez un personnage qui avance automatiquement sur un chemin tout tracé. Votre objectif est d'aller le plus loin possible en accumulant un maximum de bonus, et en esquivant au mieux les obstacles. Glissez votre doigt vers la droite ou la gauche sur l'écran pour déplacer le personnage latéralement, glissez votre doigt vers le haut pour le faire sauter, ou vers le bas pour le faire glisser.



L'image ci-dessus est une capture d'écran de l'application pendant la phase de jeu. Outre le personnage en bas de l'écran, on peut voir l'affichage de différentes informations, de gauche à droite et de haut en bas :

- le bouton de pause du jeu ;
- trois cœurs qui représentent vos vies. Si vous touchez un obstacle, vous perdez un cœur, et reprenez votre course au même endroit. Si vous perdez les trois, la phase de jeu prend fin ;
- le nombre d'arrêtes de poisson (bonus) collectées pendant la session ;
- le score (un calcul entre la distance parcourue et les bonus collectés) ;
- la distance parcourue en mètres ;
- le nombre d'objets que vous aurez croisés qui seront utilisés pour le questionnaire de mémorisation en phase 3. Tant que vous n'en aurez pas croisé au moins 7, la partie continuera (même si vous n'avez plus de cœurs). L'aspirateur rouge visible ci-dessus est un exemple de ces objets.

Le questionnaire de fin change d'une session à l'autre. Il alterne entre des questions sur votre

sentiment d'immersion, et des questions sur votre capacité à mémoriser des objets spécifiques que vous aurez croisés pendant la phase de jeu. Il arrive aussi parfois qu'aucune question ne vous soit posée.

Avant chaque session, un écran vous rappelle de brancher votre casque audio, de mettre le niveau sonore de votre téléphone au maximum, et de régler le niveau sonore du jeu à l'aide du curseur qui s'affiche.

Une compensation financière est prévue, sous forme de bons d'achat, 10€ à l'issue des 20 premières sessions accomplies, encore 10€ après les 20 suivantes, et enfin 30€ après les 30 dernières, pour un total de 50€.

La session libre

La session libre est identique à la session de test, à l'exception que vous pouvez y jouer de façon illimitée. Les informations issues des questionnaires de début et de fin (sans tâche de mémorisation toutefois) sont récoltées de la même manière. Aucune compensation financière n'est prévue pour les sessions accomplies dans ce mode.

Nous vous remercions pour votre participation.

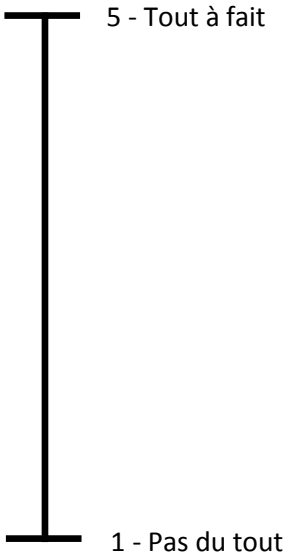
B.2 Questionnaire de débriefing

B.2.1 Questions

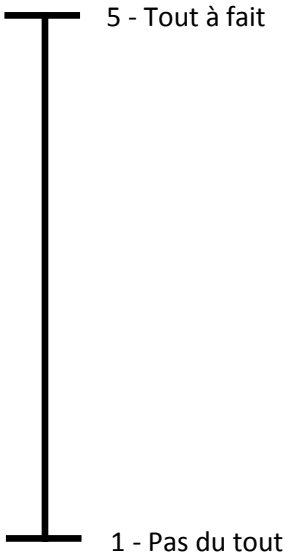
Ci-dessous, le questionnaire de fin rempli par chaque sujet à l'issue de ses sessions.

Après l'expérience :

Aimez-vous ce genre de jeu en général ?



Avez-vous apprécié le jeu ?



Avez-vous ressenti une gêne particulière à jouer dans l'environnement « forêt » (graphismes trop chargés, incohérences, bugs, etc.) ?

Avez-vous ressenti une gêne particulière à jouer dans l'environnement « ville » (graphismes trop chargés, incohérences, bugs, etc.) ?

Avez-vous ressenti une gêne particulière à jouer dans l'environnement « ville vénitienne » (graphismes trop chargés, incohérences, bugs, etc.) ?

Commentaires sur le jeu :

Commentaires sur l’expérience :

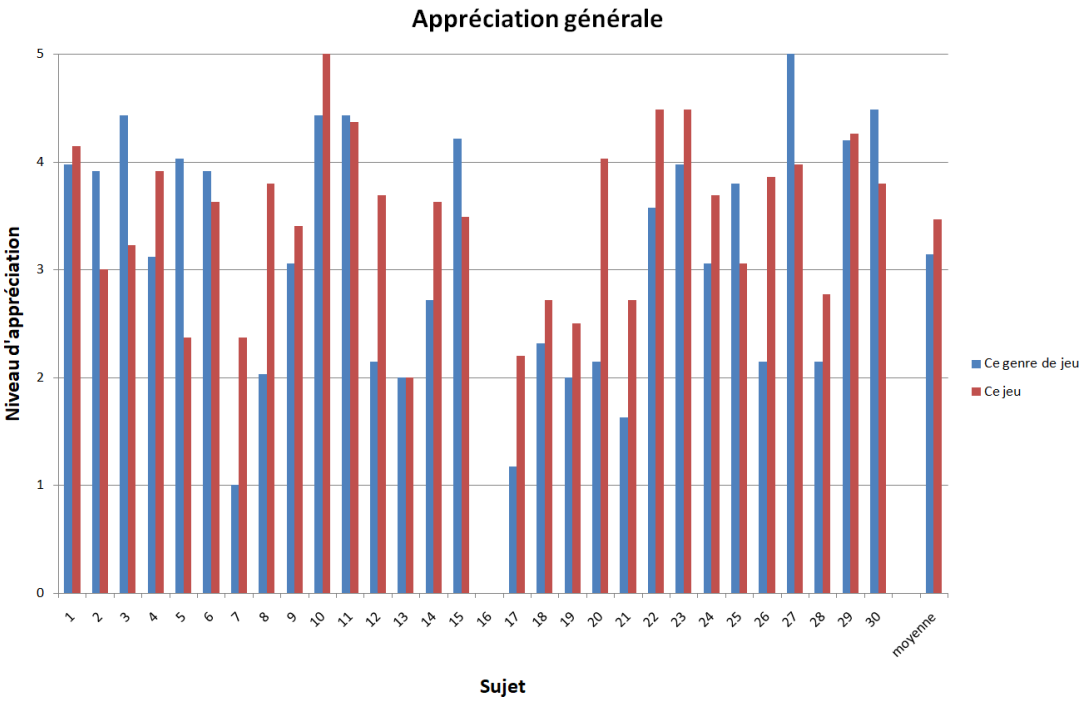


FIGURE B.1 – Représentation de l'appréciation des sujets du genre de jeu en général (en bleu) et de ce jeu en particulier (en rouge). Les données du sujet 16 n'ont pu être récoltées, pour des raisons techniques et de distance (sujet devenu injoignable après l'expérience).

B.3 Réponses

La Figure B.1 combine les résultats obtenus aux deux premières questions du questionnaire, à savoir l'appréciation générale du genre de jeu, et l'appréciation particulière de ce jeu. On constate une note moyenne meilleure pour ce jeu par rapport au genre : 18 sujets ont préféré le jeu, seulement 9 le genre, et 1 a noté les deux à l'identique. Un de nos objectifs était de proposer un jeu suffisamment abouti pour donner l'impression au sujet de jouer à un vrai jeu. Ces notes suggèrent que l'objectif est atteint. La Figure B.2 combine quant à elle les niveaux de gêne ressentie pour la forêt, la ville et la ville vénitienne. La quatrième valeur représente la gêne totale, moyenne des trois précédentes. Ici on observe une gêne moyenne plus importante ressentie pour la ville vénitienne. Cependant, le détail des notes par sujet est plus partagé : 11 sujets ont attribué une note de gêne plus importante à la ville vénitienne, contre 10 pour la forêt, 5 pour la ville et 3 au moins deux environnements à égalité. Cependant, quand elle est devant les autres, la ville vénitienne semble avoir particulièrement plus gêné. La raison tient sans doute au faut que ce niveau a été développé pour l'occasion par le prestataire Polymorph, et que son équilibrage n'a pas pu être testé dans les temps impartis. Quelques bugs reportés par les sujets dans ce niveau spécifique (notamment l'apparition parfois brusque d'un mur sur le chemin, obstruant la vue mais permettant le passage) peuvent aussi expliquer ce résultat. Une autre gêne reportée oralement est celle ressentie pendant les niveaux de nuit, en particulier dans la forêt, rendant la visibilité des obstacles moins bonne et expliquant sans doute la deuxième position de cet environnement dans les reports.

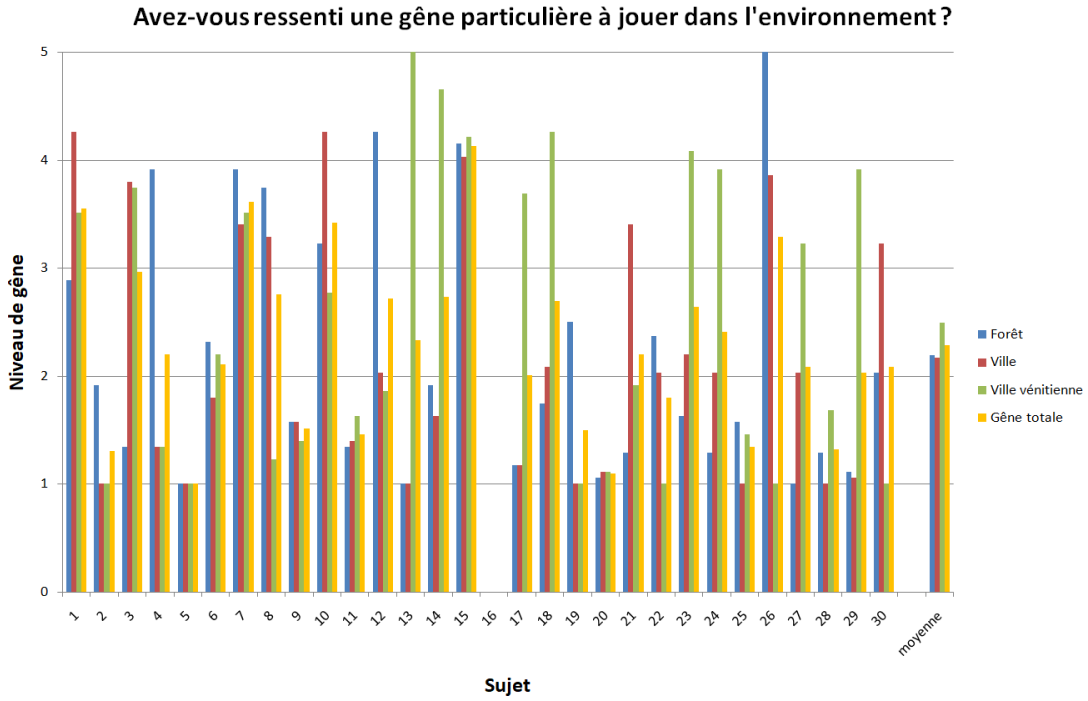


FIGURE B.2 – Représentation de la gêne des sujets ressentie pour chaque environnement. En jaune, la gêne totale, moyenne des trois précédentes. Les données du sujet 16 n'ont pu être récoltées, pour des raisons techniques et de distance (sujet devenu injoignable après l'expérience).

Bibliographie

Adomavicius, Gediminas (2011). « Context-Aware Recommender Systems ». In : *Association for the Advancement of Artificial Intelligence*, p. 1-40. issn : 0738-4602. doi : 10.1609/aimag.v32i3.2364. arXiv : 3.

Agganis, Brian T, Jeffrey A Muday et James A Schirillo (2010). « Visual biasing of auditory localization in azimuth and depth ». In : *Perceptual and Motor Skills* 111.3, p. 872-892.

Alais, David et David Burr (2004). « The ventriloquist effect results from near-optimal bimodal integration ». In : *Current biology* 14.3, p. 257-262.

Alais, David, Fiona N Newell et Pascal Mamassian (2010). « Multisensory processing in review : from physiology to behaviour ». In : *Seeing and perceiving* 23.1, p. 3-38.

Algazi, V Ralph, Richard O Duda, Ramani Duraiswami et al. (2002). « Approximating the head-related transfer function using simple geometric models of the head and torso ». In : *The Journal of the Acoustical Society of America* 112.5, p. 2053-2064.

Algazi, V Ralph, Richard O Duda, Dennis M Thompson et al. (2001). « The cipic hrtf database ». In : *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*. IEEE, p. 99-102.

André, Cédric R et al. (2014). « Subjective evaluation of the audiovisual spatial congruence in the case of stereoscopic-3D video and Wave Field Synthesis ». In : *International Journal of Human-Computer Studies* 72.1, p. 23-32.

Ang, Linus Yinn Leng, Yong Khiang Koh et Heow Pueh Lee (2017). « The performance of active noise-canceling headphones in different noise environments ». In : *Applied Acoustics* 122, p. 16-22.

Asano, Futoshi, Yoiti Suzuki et Toshio Sone (1990). « Role of spectral cues in median plane localization ». In : *The Journal of the Acoustical Society of America* 88.1, p. 159-168.

Aubert, Donatien (2019). « Les nouveaux arts de la mémoire : topiques digitales. La réactualisation de l'ars memoriae grâce à l'infographie tridimensionnelle ». Thèse de doct. Sorbonne université.

Bahu, Hélène (2016). « Localisation auditive en contexte de synthèse binaurale non-individuelle ». Thèse de doct. Université Pierre et Marie Curie-Paris VI.

Bailblé, Claude (1999). « La perception et l'attention modifiées par le dispositif cinéma ». Thèse de doct. Paris 8.

Balan, Oana et al. (2015). « The role of perceptual feedback training on sound localization accuracy in audio experiments ». In : *The International Scientific Conference eLearning and Software for Education*. T. 1. " Carol I" National Defence University, p. 502.

Barnard, Leon et al. (2007). « Capturing the effects of context on human performance in mobile computing systems ». In : *Personal and Ubiquitous Computing* 11.2, p. 81-96.

Barraclough, Nick E et al. (2005). « Integration of visual and auditory information by superior temporal sulcus neurons responsive to the sight of actions ». In : *Journal of Cognitive Neuroscience* 17.3, p. 377-391.

Bech, Søren et Geoff Martin (2005). « Attribute identification and quantification in automotive audio-part 1 : Introduction to the descriptive analysis technique ». In : *Audio Engineering Society Convention 118*. Audio Engineering Society.

Beerends, John G et Frank E De Caluwe (1999). « The influence of video quality on perceived audio quality and vice versa ». In : *Journal of the Audio Engineering Society* 47.5, p. 355-362.

Begault, Durand R, Elizabeth M Wenzel et Mark R Anderson (2001). « Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source ». In : *Journal of the Audio Engineering Society* 49.10, p. 904-916.

Bell, Andrew H et al. (2005). « Crossmodal integration in the primate superior colliculus underlying the preparation and initiation of saccadic eye movements ». In : *Journal of Neurophysiology* 93.6, p. 3659-3673.

Beresford, Kathryn et al. (2006). « Contextual effects on sound quality judgements : listening room and automotive environments ». In : *Audio Engineering Society Convention 120*. Audio Engineering Society.

Berg, Jan (2005). « OPAQUE–a tool for the elicitation and grading of audio quality attributes ». In : *Audio Engineering Society Convention 118*. Audio Engineering Society.

— (mai 2009). « The Contrasting and Conflicting Definitions of Envelopment ». In : *Audio Engineering Society Convention 126*. URL : http://www.aes.org/e-lib/browse.cfm?elib=15004.

Berg, Jan et Francis Rumsey (1999). « Spatial attribute identification and scaling by repertory grid technique and other methods ». In : *Proceedings of the AES 16th International Conference on Spatial Sound Reproduction*.

— (2000). « In search of the spatial dimensions of reproduced sound : Verbal protocol analysis and cluster analysis of scaled verbal descriptors ». In : *Audio Engineering Society Convention 108*. Audio Engineering Society.

— (2001). « Verification and correlation of attributes used for describing the spatial quality of reproduced sound. » In :

— (2002). « Validity of selected spatial attributes in the evaluation of 5-channel microphone techniques ». In : *AES Convention : 10/05/2002-13/05/2002*. Audio Engineering Society, Inc.

Bertelson, Paul et Gisa Aschersleben (1998). « Automatic visual bias of perceived auditory location ». In : *Psychonomic bulletin & review* 5.3, p. 482-489.

Bertelson, Paul et Monique Radeau (1976). « Ventriloquism, sensory interaction, and response bias : Remarks on the paper by Choe, Welch, Gilford, and Juola ». In : *Perception & Psychophysics* 19.6, p. 531-535.

— (1981). « Cross-modal bias and perceptual fusion with auditory-visual spatial discordance ». In : *Perception & psychophysics* 29.6, p. 578-584.

Bertelson, Paul, Jean Vroomen et al. (2000). « The ventriloquist effect does not depend on the direction of deliberate visual attention ». In : *Perception & psychophysics* 62.2, p. 321-332.

Beyer, Justus et Sebastian Möller (2014a). « Assessing the Impact of Game Type, Display Size and Network Delay on Mobile Gaming QoE ». In : *PIK-Praxis der Informationsverarbeitung und Kommunikation* 37.4, p. 287-295.

— (2014b). « Gaming ». In : *Quality of experience*. Springer, p. 367-381.

Blauert, Jens (1997). *Spatial hearing : the psychophysics of human sound localization*. MIT press.

— (2013). *The technology of binaural listening*. Springer.

Blauert, Jens et Ute Jekosch (2003). « Concepts behind sound quality : Some basic considerations ». In : *Proceedings of Inter-Noise 2003*, p. 72-79.

— (2012). « A layer model of sound quality ». In : *Journal of the Audio Engineering Society* 60.1/2, p. 4-12.

Blumlein, Alan D (1958). « British patent specification 394,325 (improvements in and relating to sound-transmission, sound-recording and sound-reproducing systems) ». In : *Journal of the Audio Engineering Society* 6.2, p. 91-130.

Böhmer, Matthias et al. (2011). « Falling asleep with Angry Birds, Facebook and Kindle : a large scale study on mobile application usage ». In : *Proceedings of the 13th international conference on Human computer interaction with mobile devices and services*. ACM, p. 47-56.

Bolger, Niall, Angelina Davis et Eshkol Rafaeli (2003). « Diary methods : Capturing life as it is lived ». In : *Annual review of psychology* 54.1, p. 579-616.

Bowen, Amanda L et al. (2011). « Visual signals bias auditory targets in azimuth and depth ». In : *Experimental brain research* 214.3, p. 403-414.

Bradley, Margaret M et Peter J Lang (1994). « Measuring emotion : the self-assessment manikin and the semantic differential ». In : *Journal of behavior therapy and experimental psychiatry* 25.1, p. 49-59.

Brockmann, Tobias, Stefan Stieglitz et Christoph Lattemann (2014). « Taxonomy of enterprise-related mobile applications ». In : *International Conference on Social Computing and Social Media*. Springer, p. 37-47.

Brockmyer, Jeanne H et al. (2009). « The development of the Game Engagement Questionnaire : A measure of engagement in video game-playing ». In : *Journal of Experimental Social Psychology* 45.4, p. 624-634.

Bronkhorst, Adelbert W (1995). « Localization of real and virtual sound sources ». In : *The Journal of the Acoustical Society of America* 98.5, p. 2542-2553.

Brooks, Peter et Bjoørn Hestnes (2010). « User measures of quality of experience : Why being objective and quantitative is important ». In : *IEEE Network* 24.2, p. 8-13. issn : 08908044. doi : 10.1109/MNET.2010.5430138.

Brown, Barry, Moira McGregor et Eric Laurier (2013). « iPhone in vivo : video analysis of mobile device use ». In : *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM, p. 1031-1040.

Bruijn, Werner PJ de et Marinus M Boone (2003). « Application of Wave Field Synthesis in life-size videoconferencing ». In : *Audio Engineering Society Convention 114*. Audio Engineering Society.

Buchinger, Shelley et al. (2010). « Slider or glove ? Proposing an alternative quality rating methodology ». In :

Burr, David et David Alais (2006). « Combining visual and auditory information ». In : *Progress in brain research* 155, p. 243-258.

Carpentier, Thibaut et al. (2014). « Measurement of a head-related transfer function database with high spatial resolution ». In : *7th Forum Acusticum (EAA)*.

Catellier, Andrew et al. (2012). « Impact of mobile devices and usage location on perceived multimedia quality ». In : *2012 Fourth International Workshop on Quality of Multimedia Experience*. IEEE, p. 39-44.

Chen, Lihan et Jean Vroomen (2013). « Intersensory binding across space and time : a tutorial review ». In : *Attention, Perception, & Psychophysics* 75.5, p. 790-811.

Cheon, Manri et al. (2015). « Quality assessment of mobile videos ». In : *Visual Signal Quality Assessment*. Springer, p. 99-127.

Chion, Michel (2013). *L'audio-vision : son et image au cinéma*. Armand Colin.

Choe, Chong S et al. (1975). « The "ventriloquist effect" : Visual dominance or response bias ? » In : *Perception & Psychophysics* 18.1, p. 55-60.

Choisel, Sylvain et Florian Wickelmaier (2006). « Extraction of Auditory Features and Elicitation of Attributes for the Assessment of Multichannel Reproduced Sound ». In : *J. Audio Eng. Soc* 54.9, p. 815-826. URL : http://www.aes.org/e-lib/browse.cfm?elib=13903.

Cobos, Maximo et al. (2015). « Subjective quality assessment of multichannel audio accompanied with video in representative broadcasting genres ». In : *Multimedia Systems* 21.4, p. 363-379.

Consolvo, Sunny et Miriam Walker (2003). « Using the experience sampling method to evaluate ubicomp applications ». In : *IEEE Pervasive Computing* 2.2, p. 24-31.

Csikszentmihalyi, Mihaly, Sami Abuhamdeh et Jeanne Nakamura (2014). « Flow ». In : *Flow and the foundations of positive psychology*. Springer, p. 227-238.

Cyan World (1993). *Myst*. [Mac, then others].

De Masi, Alexandre et al. (2016). « mQoL smart lab : quality of life living lab for interdisciplinary experiments ». In : *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing : Adjunct*. ACM, p. 635-640.

De Moor, Katrien (2012). « Are engineers from Mars and users from Venus ? : bridging gaps in quality of experience research : reflections on and experiences from an interdisciplinary journey ». eng. Thèse de doct. Ghent University, p. XXI, 342.

De Moor, Katrien et al. (2010). « Proposed framework for evaluating quality of experience in a mobile, testbed-oriented living lab setting ». In : *Mobile Networks and Applications* 15.3, p. 378-391.

De Pessemier, Toon et al. (2012). « Quantifying subjective quality evaluations for mobile video watching in a semi-living lab context ». In : *IEEE Transactions on Broadcasting* 58.4, p. 580-589.

Dey, Anind K, Gregory D Abowd et Daniel Salber (2001). « A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications ». In : *Human–Computer Interaction* 16.2-4, p. 97-166.

DOWINO (2015). *A Blind Legend*. [Android, iOS, Steam].

Dubois, Danièle (2000). « Categories as acts of meaning : The case of categories in olfaction and audition ». In : *Cognitive science quarterly* 1.1, p. 35-68.

Egger, Sebastian, Peter Reichl et Katrin Schoenenberg (2014). « Quality of Experience and Interactivity ». In : *Quality of Experience*. Springer, p. 149-161.

Eisler, Hannes (1966). « Measurement of Perceived Acoustic Quality of Sound-Reproducing Systems by Means of Factor Analysis ». In : *The Journal of the Acoustical Society of America* 39.3, p. 484-492.

Engelke, Ulrich et al. (2016). « Psychophysiology-Based QoE Assessment : A Survey ». In : *IEEE Journal of Selected Topics in Signal Processing*.

Ernst, Marc O et Martin S Banks (2002). « Humans integrate visual and haptic information in a statistically optimal fashion ». In : *Nature* 415.6870, p. 429.

Ernst, Marc O et Heinrich H Bülthoff (2004). « Merging the senses into a robust percept ». In : *Trends in cognitive sciences* 8.4, p. 162-169.

Ewert, P Harry (1930). « A study of the effect of inverted retinal stimulation upon spatially coordinated behavior. » In : *Genetic Psychology Monographs*.

Farnell, Andy (2010). *Designing sound*. Mit Press.

Faure, Julien et Grégory PALLONE (2005). *Evaluation de la synthese binaurale dynamique*. Rapp. tech. Tech. Rep., France Telecom.

Fiebig, André (2015). « Influence of context effects on sound quality assessments ». In : *Proceedings of EuroNoise*, p. 2555-2560.

Følstad, Asbjørn (2008). « Towards a living lab for development of online community services ». In :

Francombe, Jon, Timothy Brookes et Russell Mason (2015). « Perceptual Evaluation of Spatial Audio : Where Next ? » In : *Proceedings of the 22nd International Congress on Sound and Vibration, Florence, Italy, 12–16 July*.

— (2017). « Evaluation of spatial audio reproduction methods (part 1) : elicitation of perceptual differences ». In : *Journal of the Audio Engineering Society* 65.3, p. 198-211.

Francombe, Jon, Timothy Brookes, Russell Mason et James Woodcock (2016). « Determining and labeling the preference dimensions of spatial audio replay ». In : *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, p. 1-6.

— (2017). « Evaluation of spatial audio reproduction methods (part 2) : analysis of listener preference ». In : *Journal of the Audio Engineering Society* 65.3, p. 212-225.

Frank, Matthias et al. (2014). « Spatial audio rendering ». In : *Quality of Experience*. Springer, p. 247-260.

Froehlich, Jon et al. (2007). « MyExperience : a system for in situ tracing and capturing of user feedback on mobile phones ». In : *Proceedings of the 5th international conference on Mobile systems, applications and services*. ACM, p. 57-70.

Gabrielsson, Alf (1979). « Dimension analyses of perceived sound quality of sound-reproducing systems ». In : *Scandinavian Journal of Psychology* 20.1, p. 159-169.

Gabrielsson, Alf et Håkan Sjögren (1979). « Perceived sound quality of sound-reproducing systems ». In : *The Journal of the Acoustical Society of America* 65.4, p. 1019-1033. doi : 10.1121/1.382579.

Gardner, Mark B (1968). « Proximity image effect in sound localization ». In : *The Journal of the Acoustical Society of America* 43.1, p. 163-163.

Geier, Matthias et al. (2010). « Perceptual evaluation of focused sources in wave field synthesis ». In : *Audio Engineering Society Convention 128*. Audio Engineering Society.

Giacalone, Davide et al. (mai 2017). « Sensory Profiling of High-End Loudspeakers Using Rapid Methods—Part 2 : Projective Mapping with Expert and Naïve Assessors ». In : *Audio Engineering Society Convention 142*. url : http://www.aes.org/e-lib/browse.cfm?elib=18651.

Gibbs, Chris, Ulrike Gretzel et Jesse Saltzman (2016). « An experience-based taxonomy of branded hotel mobile application features ». In : *Information Technology & Tourism* 16.2, p. 175-199.

Golder, Scott A et Michael W Macy (2011). « Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures ». In : *Science* 333.6051, p. 1878-1881.

Gonot, Antoine, Marc Emerit et Noël Château (2006). « Usability of 3D-sound for navigation in a constrained virtual environment ». In : *AES, 120th Audio Engineering Society Convention, Paris*. Audio Engineering Society.

Gonzalez, Marta C, Cesar A Hidalgo et Albert-Laszlo Barabasi (2008). « Understanding individual human mobility patterns ». In : *nature* 453.7196, p. 779.

Grani, Francesco et al. (2014). « Design and evaluation of Binaural auditory rendering for CAVEs ». In : *2014 IEEE Virtual Reality (VR)*. IEEE, p. 73-74.

Guastavino, Catherine (2003). « Étude sémantique et acoustique de la perception des basses fréquences dans l'environnemet sonore urbain ». Thèse de doct. Université Paris 6.

Guastavino, Catherine et Pascale Chemiliée (2003). « Conceptualisations en langue, représentations cognitives et validité écologique : une approche psycholinguistique de la perception des basses fréquences. » In : *Psychologie française*.

Guastavino, Catherine et Brian FG Katz (2004). « Perceptual evaluation of multi-dimensional spatial audio reproduction ». In : *The Journal of the Acoustical Society of America* 116.2, p. 1105-1115.

Gueguen, Marc (2011). « Intégration multisensorielle et variabilité interindividuelle ». Thèse de doct. Université Paris Sud-Paris XI.

Guezenoc, Corentin et Renaud Seguier (2018). « HRTF Individualization : A Survey ». In : *Audio Engineering Society Convention 145*. Audio Engineering Society.

Halfbrick Studios (2011). *Jetpack Joyride*. [iOS].

Hamasaki, Kimio, Yasushige Nakayama et al. (2007). « Wide listening area with exceptional spatial sound quality of a 22.2 multichannel sound system ». In : *Audio Engineering Society Convention 122*. Audio Engineering Society.

Hamasaki, Kimio, Toshiyuki Nishiguchi et al. (2004). « Advanced multichannel audio systems with superior impression of presence and reality ». In : *Audio Engineering Society Convention 116*. Audio Engineering Society.

Hammer, Florian, Sebastian Egger-Lampl et Sebastian Möller (juil. 2018). « Quality-of-user-experience : a position paper ». In : *Quality and User Experience* 3.1, p. 9. issn : 2366-0147. doi : 10.1007/s41233-018-0022-0. url : https://doi.org/10.1007/s41233-018-0022-0.

Hassenzahl, Mare et al. (2000). « Hedonic and ergonomic quality aspects determine a software's appeal ». In : *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, p. 201-208.

Hendrickx, Etienne et al. (2015). « Ventriloquism effect with sound stimuli varying in both azimuth and elevation ». In : *The Journal of the Acoustical Society of America* 138.6, p. 3686-3697.

Henze, Niels, Enrico Rukzio et Susanne Boll (2011). « 100,000,000 taps : analysis and improvement of touch performance in the large ». In : *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*. ACM, p. 133-142.

Hettinger, Lawrence J et Gary E Riccio (1992). « Visually induced motion sickness in virtual environments ». In : *Presence : Teleoperators & Virtual Environments* 1.3, p. 306-310.

Hládek, L'uboš et al. (2013). « Ventriloquism effect and aftereffect in the distance dimension ». In : *Proceedings of Meetings on Acoustics ICA2013*. T. 19. 1. ASA, p. 050042.

Hofman, Paul M, Jos GA Van Riswick et A John Van Opstal (1998). « Relearning sound localization with new ears ». In : *Nature neuroscience* 1.5, p. 417.

Hollier, Mike P et al. (1999). « Multi-modal perception ». In : *BT Technology Journal* 17.1, p. 35-46.

Honda, Akio, Hiroshi Shibata, Jiro Gyoba et al. (2007). « Transfer effects on sound localization performances from playing a virtual three-dimensional auditory game ». In : *Applied Acoustics* 68.8, p. 885-896.

Honda, Akio, Hiroshi Shibata, Souta Hidaka et al. (2013). « Effects of head movement and proprioceptive feedback in training of sound localization ». In : *i-Perception* 4.4, p. 253-264.

Hong, Qi et Wen-Ming Zhu (2017). « Research on HAS video QoE assessment technology ». In : *International Journal of Wireless and Mobile Computing* 13.3, p. 245-252.

Hossfeld, Tobias, Florian Metzger et Michael Jarschel (2015). « Qoe for cloud gaming ». In : *E-LETTER*.

Howard, Ian P et William B Templeton (1966). *Human spatial orientation.* John Wiley & Sons.

Huizinga, Johan et Cécile Seresia (1952). « Homo ludens, essai sur la fonction sociale du jeu ». In :

Ickin, Selim et al. (2012). « Factors influencing quality of experience of commonly used mobile applications ». In : *IEEE Communications Magazine* 50.4, p. 48-56.

Imangi Studios (2011). *Temple Run.* [iOS].

ISO (1994). *Quality management and quality assurance standards (ISO 8402).* International Organization for Standardization.

— (2000). *Quality Management Systems–Fundamentals and Vocabulary (ISO 9000).* International Organization for Standardization.

— (2015). *Quality Management Systems–Fundamentals and Vocabulary (ISO 9000).* International Organization for Standardization.

ITU-R (2002). « BT.500-11 : Methodology for the subjective assessment of the quality of television pictures ». In :

— (2012). « BS775-3 : Multichannel stereophonic sound system with and without accompanying picture ». In :

— (2014). « BS.2300-0 : Methods for assessor screening ». In :

— (2015a). « BS.1116-3 : Methods for the subjective assessment of small impairments in audio systems ». In :

— (2015b). « BS.1534-3 : Method for the subjective assessment of intermediate quality level of audio systems ». In :

— (2018). « BS.2051-2 : Advanced sound system for programme production ». In :

— (2019). « BS.1284-2 : General methods for the subjective assessment of sound quality ». In :

ITU-T (1996). « P.800 : Methods for subjective determination of transmission quality ». In :

— (1998). « P.911 : Subjective audiovisual quality assessment methods for multimedia applications ». In :

— (2008). « G.1080 : Quality of experience requirements for IPTV services ». In :

— (2014). « G.1091 - Quality of Experience requirements for telepresence services ». In :

— (2015). « G.107 : The E-model - a computational model for use in transmission planning ». In :

— (2017). « P.10/G.100 : Vocabulary for performance, quality of service and quality of experience ». In :

Jack, Charles E et Willard R Thurlow (1973). « Effects of degree of visual association and angle of displacement on the "ventriloquism" effect. » In : *Perceptual and motor skills.*

Jackson, CV (1953). « Visual factors in auditory localization ». In : *Quarterly Journal of Experimental Psychology* 5.2, p. 52-65.

Jacucci, Giulio et al. (2007). « Comedia : mobile group media for active spectatorship ». In : *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* ACM, p. 1273-1282.

Jarschel, Michael et al. (2011). « An evaluation of QoE in cloud gaming based on subjective tests ». In : *2011 Fifth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*. IEEE, p. 330-335.

Jekosch, Ute (2006). *Voice and speech quality perception : assessment and evaluation*. Springer Science & Business Media.

Jennett, Charlene et al. (2008). « Measuring and defining the experience of immersion in games ». In : *International journal of human-computer studies* 66.9, p. 641-661.

Jumisko-Pyykkö, Satu (2008). « "I would like to see the subtitles and the face or at least hear the voice" : Effects of picture ratio and audio–video bitrate ratio on perception of quality in mobile television ». In : *Multimedia Tools and Applications* 36.1-2, p. 167-184.

— (2011). « User-centered quality of experience and its evaluation methods for mobile television ». Thèse de doct. Tampere University of Technology.

Jumisko-Pyykkö, Satu et Jukka Häkkinen (2008). « Profiles of the evaluators : impact of psychographic variables on the consumer-oriented quality assessment of mobile television ». In : *Multimedia on Mobile Devices 2008*. T. 6821. International Society for Optics et Photonics, p. 68210L.

Jumisko-Pyykkö, Satu et Miska M Hannuksela (2008). « Does context matter in quality evaluation of mobile television ? » In : *Proceedings of the 10th international conference on Human computer interaction with mobile devices and services*. ACM, p. 63-72.

Jumisko-Pyykkö, Satu, Dominik Strohmeier et al. (2010). « Descriptive quality of experience for mobile 3D video ». In : *Proceedings of the 6th Nordic Conference on Human-Computer Interaction : Extending Boundaries*. ACM, p. 266-275.

Jumisko-Pyykkö, Satu et Timo Utriainen (2011). « A hybrid method for quality evaluation in the context of use for mobile (3D) television ». In : *Multimedia Tools and Applications* 55.2, p. 185-225.

Jumisko-Pyykkö, Satu et Teija Vainio (2010). « Framing the context of use for mobile HCI ». In : *International journal of mobile human computer interaction (IJMHCI)* 2.4, p. 1-28.

Kaikkonen, Anne et al. (2005). « Usability testing of mobile applications : A comparison between laboratory and field testing ». In : *Journal of Usability studies* 1.1, p. 4-16.

Katz, Brian FG et Rozenn Nicol (2018). « Binaural spatial reproduction ». In : *Sensory Evaluation of Sound*. CRC Press.

Kennedy-Eden, Heather et Ulrike Gretzel (2012). « A taxonomy of mobile applications in tourism ». In :

Kersten, Daniel, Pascal Mamassian et Alan Yuille (2004). « Object perception as Bayesian inference ». In : *Annu. Rev. Psychol.* 55, p. 271-304.

Khan, Asiya, Lingfen Sun et Emmanuel Ifeachor (2012). « QoE prediction model and its application in video quality adaptation over UMTS networks ». In : *IEEE Transactions on Multimedia* 14.2, p. 431-442.

Kiloo & Sybo Games (2012). *Subway Surfers*. [iOS, Android and others].

Kim, Sang-Myeong et Wonjae Choi (2005). « On the externalization of virtual sound images in headphone reproduction : A Wiener filter approach ». In : *The Journal of the Acoustical Society of America* 117.6, p. 3657-3665.

King, AJ et AR Palmer (1985). « Integration of visual and auditory information in bimodal neurones in the guinea-pig superior colliculus ». In : *Experimental Brain Research* 60.3, p. 492-500.

Kjeldsen, Annie D (1998). « The measurement of personal preferences by repertory grid technique ». In : *Audio Engineering Society Convention 104*. Audio Engineering Society.

Komiyama, Setsu (1989). « Subjective evaluation of angular displacement between picture and sound directions for HDTV sound systems ». In : *Journal of the Audio Engineering Society* 37.4, p. 210-214.

Korhonen, Hannu, Juha Arrasvuori et Kaisa Väänänen-Vainio-Mattila (2010). « Analysing user experience of personal mobile products through contextual factors ». In : *Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia*. ACM, p. 11.

Langendijk, Erno HA et Adelbert W Bronkhorst (2000). « Fidelity of three-dimensional-sound reproduction using a virtual auditory display ». In : *The Journal of the Acoustical Society of America* 107.1, p. 528-537.

— (2002). « Contribution of spectral cues to human sound localization ». In : *The Journal of the Acoustical Society of America* 112.4, p. 1583-1596.

Larson, Reed et Mihaly Csikszentmihalyi (2014). « The experience sampling method ». In : *Flow and the foundations of positive psychology*. Springer, p. 21-34.

Larsson, Pontus, Daniel Vastfjall et Mendel Kleiner (2002). « Better presence and performance in virtual environments by improved binaural sound rendering ». In : *Audio Engineering Society Conference : 22nd International Conference : Virtual, Synthetic, and Entertainment Audio*. Audio Engineering Society.

Lassalle, Julie (2013). « Etude de l'influence de la qualité audiovisuelle sur la qualité d'expérience du spectateur : combinaison d'indicateurs subjectifs, physiologiques et oculaires ». Thèse de doct. Télécom Bretagne, Université de Bretagne-Sud.

Lassalle, Julie, Laetitia Gros et Gilles Coppin (2011). « Combination of physiological and subjective measures to assess quality of experience for audiovisual technologies ». In : *2011 Third international workshop on quality of multimedia experience*. IEEE, p. 13-18.

Le Bagousse, Sarah, Mathieu Paquier et Catherine Colomes (avr. 2012). « Assessment of spatial audio quality based on sound attributes ». In : *Acoustics 2012*. Nantes, France, p. 873-877. URL : https://hal.univ-brest.fr/hal-00827941.

— (2014). « Categorization of Sound Attributes for Audio Quality Assessment—A Lexical Study ». In : *J. Audio Eng. Soc* 62.11, p. 736-747. URL : http://www.aes.org/e-lib/browse.cfm?elib=17549.

Le Bagousse, Sarah, Mathieu Paquier, Catherine Colomes et Samuel Moulin (oct. 2011). « Sound Quality Evaluation Based on Attributes - Application to Binaural Contents ». In : *Audio Engineering Society Convention 131*. URL : http://www.aes.org/e-lib/browse.cfm?elib=16068.

Le Callet, Patrick et al. (2013). « Qualinet White Paper on Definitions of Quality of Experience, version 1.2 ». In :

Lecomte, Pierre et al. (2015). « On the use of a Lebedev grid for Ambisonics ». In : *Audio Engineering Society Convention 139*. Audio Engineering Society.

Lee, Sojeong, Hwayeong Kang et Gwanseob Shin (2015). « Head flexion angle while using a smartphone ». In : *Ergonomics* 58.2, p. 220-226.

Lee, Yeng-Ting et al. (2012). « Are all games equally cloud-gaming-friendly ? : an electromyographic approach ». In : *Proceedings of the 11th annual workshop on network and systems support for games*. IEEE Press, p. 3.

Letowski, T (1989). « Sound quality assessment : concepts and criteria ». In : *87th Convention of the Audio Engineering Society, J. Audio Eng. Soc.(Abstracts)*. T. 37, p. 1062.

Lewald, Jörg, Walter H Ehrenstein et Rainer Guski (2001). « Spatio-temporal constraints for auditory–visual integration ». In : *Behavioural brain research* 121.1, p. 69-79.

Lewald, Jörg et Rainer Guski (2003). « Cross-modal perceptual integration of spatially and temporally disparate auditory and visual stimuli ». In : *Cognitive brain research* 16.3, p. 468-478.

Li, En et Donnel Briley (2011). « Attitudes shaped by eye movements : The reading direction effect ». In : *ACR North American Advances*.

Li, Jing et al. (2018). « Quantifying the Influence of Devices on Quality of Experience for Video Streaming ». In : *2018 Picture Coding Symposium (PCS)*. IEEE, p. 308-312.

Liang, Ting-Peng et Yi-Hsuan Yeh (2011). « Effect of use contexts on the continuous use of mobile services : the case of mobile games ». In : *Personal and Ubiquitous Computing* 15.2, p. 187-196.

Lindau, Alexander et al. (2014). « A Spatial Audio Quality Inventory (SAQI) ». In : *Acta Acustica united with Acustica* 100.5, p. 984-994. issn : 1610-1928. doi : doi:10.3813/AAA.918778.

Liu, Ning, Ying Liu et Xia Wang (2010). « Data logging plus e-diary : towards an online evaluation approach of mobile service field trial ». In : *Proceedings of the 12th international conference on Human computer interaction with mobile devices and services*. ACM, p. 287-290.

Liu, Tao et al. (2012). « Continuous mobile video subjective quality assessment using gaming steering wheel ». In : *Sixth International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM). Scottsdale, Arizona*.

Lorho, Gaëtan (mai 2005a). « Evaluation of Spatial Enhancement Systems for Stereo Headphone Reproduction by Preference and Attribute Rating ». In : *Audio Engineering Society Convention 118*. url : http://www.aes.org/e-lib/browse.cfm?elib=13230.

— (2005b). « Individual vocabulary profiling of spatial enhancement systems for stereo headphone reproduction ». In : *Audio Engineering Society Convention 119*. Audio Engineering Society.

Majdak, Piotr et al. (2013). « Spatially oriented format for acoustics : A data exchange format representing head-related transfer functions ». In : *Audio Engineering Society Convention 134*. Audio Engineering Society.

Mannerheim, Paul (2011). « Spatial sound and stereoscopic vision ». In : *Audio Engineering Society Convention 130*. Audio Engineering Society.

Martens, Harald et Magni Martens (2001). *Multivariate analysis of quality : an introduction*. John Wiley & Sons.

Martens, William (2001). « Uses and misuses of psychophysical methods in the evaluation of spatial sound reproduction ». In : *Audio Engineering Society Convention 110*. Audio Engineering Society.

Martens, William L. et Nick Zacharov (sept. 2000). « Multidimensional Perceptual Unfolding of Spatially Processed Speech I : Deriving Stimulus Space Using INDSCAL ». In : *Audio Engineering Society Convention 109*. url : http://www.aes.org/e-lib/browse.cfm?elib=9114.

Martin, Russell L, Ken I McAnally et Melis A Senova (2001). « Free-field equivalent localization of virtual audio ». In : *Journal of the Audio Engineering Society* 49.1/2, p. 14-22.

Mattila, Ville-Veikko (nov. 2001). « Descriptive Analysis of Speech Quality in Mobile Communications : Descriptive Language Development and External Preference

Mapping ». In : *Audio Engineering Society Convention 111*. URL : http://www.aes.org/e-lib/browse.cfm?elib=9880.

Mausfeld, Rainer (2003). « Conjoint representations and the mental capacity for multiple simultaneous perspectives ». In :

Mauss, Iris B et al. (2005). « The tie that binds ? Coherence among emotion experience, behavior, and physiology. » In : *Emotion* 5.2, p. 175.

McDermott, Barbara J (1969). « Multidimensional analyses of circuit quality judgments ». In : *The Journal of the Acoustical Society of America* 45.3, p. 774-781.

McGurk, Harry et John MacDonald (1976). « Hearing lips and seeing voices ». In : *Nature* 264, p. 746-748.

McMullin, Elisabeth et al. (2018). « Developing a Method for the Subjective Evaluation of Smartphone Music Playback ». In : *Audio Engineering Society Convention 145*. Audio Engineering Society.

Melchior, Frank, Sandra Brix et al. (2003). « Wave field syntheses in combination with 2d video projection ». In : *Audio Engineering Society Conference : 24th International Conference : Multichannel Audio, The New Reality*. Audio Engineering Society.

Melchior, Frank, Jens-Oliver Fischer et Diemer de Vries (2006). « Audiovisual perception using wave field synthesis in combination with augmented reality systems : Horizontal positioning ». In : *AES 28th International Conference*, p. 1-10.

Menkovski, Vlado et al. (2009). « Predicting quality of experience in multimedia streaming ». In : *Proceedings of the 7th International Conference on Advances in Mobile Computing and Multimedia*. ACM, p. 52-59.

Meredith, M Alex, James W Nemitz et Barry E Stein (1987). « Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors ». In : *Journal of Neuroscience* 7.10, p. 3215-3229.

Meredith, M Alex et Barry E Stein (1983). « Interactions among converging sensory inputs in the superior colliculus ». In : *Science* 221.4608, p. 389-391.

Mershon, Donald H et al. (1980). « Visual capture in auditory distance perception : Proximity image effect reconsidered. » In : *Journal of Auditory Research*.

Michaud, Pierre-Yohan et al. (2013). « Perceptual evaluation of dissimilarity between auditory stimuli : an alternative to the paired comparison ». In : *Acta Acustica united with Acustica* 99.5, p. 806-815.

Miller, George A (1956). « The magical number seven, plus or minus two : Some limits on our capacity for processing information. » In : *Psychological review* 63.2, p. 81.

Minnaar, Pauli et al. (2001). « Localization with binaural recordings from artificial and human heads ». In : *Journal of the Audio Engineering Society* 49.5, p. 323-336.

Mitra, Karan, Arkady Zaslavsky et Christer Ahlund (2015). « Context-Aware QoE Modelling, Measurement, and Prediction in Mobile Computing Systems ». In : *IEEE Transactions on Mobile Computing* 14.5, p. 920-936. ISSN : 15361233. DOI : 10.1109/TMC.2013.155.

Moeck, Thomas et al. (2007). « Progressive perceptual audio rendering of complex scenes ». In : *Proceedings of the 2007 symposium on Interactive 3D graphics and games*. ACM, p. 189-196.

Möller, Andreas et al. (2013). « Investigating self-reporting behavior in long-term studies ». In : *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, p. 2931-2940.

Møller, Henrik, Dorte Hammershøi et al. (1999). « Evaluation of artificial heads in listening tests ». In : *Journal of the Audio Engineering Society* 47.3, p. 83-100.

MØLLER, Henrik, Clemen Boje JENSEN et al. (1996). « Using a typical human subject for binaural recording ». In : *Audio Engineering Society Convention 100*. Audio Engineering Society.

MØLLER, Henrik, Michael Friis SØRENSEN, Dorte HAMMERSHØI et al. (1995). « Head-related transfer functions of human subjects ». In : *Journal of the Audio Engineering Society* 43.5, p. 300-321.

MØLLER, Henrik, Michael Friis SØRENSEN, Clemen Boje JENSEN et al. (1996). « Binaural technique : Do we need individual recordings ? » In : *Journal of the Audio Engineering Society* 44.6, p. 451-469.

MÖLLER, Sebastian, Dennis POMMER et al. (2013). « Factors influencing gaming QoE : Lessons learned from the evaluation of cloud gaming services ». In : *Proceedings of the 4th International Workshop on Perceptual Quality of Systems (PQS 2013)*, p. 1-5.

MÖLLER, Sebastian et Alexander RAAKE (2014). « Quality of Experience : Terminology, methods and applications ». In : *PIK-Praxis der Informationsverarbeitung und Kommunikation* 37.4, p. 255-263.

MÖLLER, Sebastian, Marcel WÄLTERMANN et Marie-Neige GARCIA (2014). « Features of quality of experience ». In : *Quality of Experience*. Springer, p. 73-84.

MOLLER, Sebastian et al. (2009). « A taxonomy of quality of service and quality of experience of multimodal human-machine interaction ». In : *2009 International Workshop on Quality of Multimedia Experience*. IEEE, p. 7-12.

MOULIN, Samuel, Rozenn NICOL et Laetitia GROS (2012). « Spatial audio quality in regard to 3D video ». In : *Acoustics 2012*.

MRSIC-FLOGEL, Thomas D et al. (2001). « Listening through different ears alters spatial response fields in ferret primary auditory cortex ». In : *Journal of Neurophysiology* 86.2, p. 1043-1046.

NAGEL, Frederik et al. (2007). « EMuJoy : Software for continuous measurement of perceived emotions in music ». In : *Behavior Research Methods* 39.2, p. 283-290.

NAKAYAMA, Takeshi et al. (1971). « Subjective assessment of multichannel reproduction ». In : *Journal of the Audio Engineering Society* 19.9, p. 744-751.

NGUYEN, Khoa-Van (2012). « Technologie binaurale et contexte de réalité virtuelle : études perceptives et optimisation ». Thèse de doct. Université Pierre et Marie Curie-Paris 6.

NGUYEN, Khoa-Van et al. (2010). « Etudes de l'intégration visuo-auditive dans un contexte de réalité virtuelle. » In : *10ème Congrès Français d'Acoustique*.

NICKERSON, Robert et al. (2009). « Taxonomy development in information systems : Developing a taxonomy of mobile applications ». In : *European conference in information systems*, p. xxx-xxx.

NICOL, Rozenn, Marc EMERIT et Laetitia GROS (2018). « HRTF prêt-a-porter pour le son binaural dans les futurs contenus d'Orange ». In : *Actes du 14ème Congrès Français d'Acoustique*.

NICOL, Rozenn, Marc EMERIT, Edwige RONCIÈRE et al. (2016). « How to make immersive audio available for mass-market listening ». In : *EBU Technical Review* 14.

NICOL, Rozenn, Laetitia GROS, Catherine COLOMES, Markus NOISTERNIG et al. (2014). « A roadmap for assessing the quality of experience of 3D audio binaural rendering ». In : *Proceedings of the EAA Joint Symposium on Auralization and Ambisonics 2014*. Universitätsverlag der TU Berlin, p. 100-106.

NICOL, Rozenn, Laetitia GROS, Catherine COLOMES, E RONCIÈRE et al. (2016). « Etude comparative du rendu de différentes techniques de prise de son spatialisée après binauralisation ». In : *CFA/VISHNO*.

Pappas, Ilias O et al. (2019). « Explaining user experience in mobile gaming applications : an fsQCA approach ». In : *Internet Research*.

Parseihian, Gaëtan et Brian FG Katz (2012). « Rapid head-related transfer function adaptation using a virtual auditory environment ». In : *The Journal of the Acoustical Society of America* 131.4, p. 2948-2957.

Pedersen, Torben Holm (mar. 2005). « The semantic shape of sounds, Lexicon of sound-describing words ». In :

Pedersen, Torben Holm et Nick Zacharov (2008). « How many psycho-acoustic attributes are needed ». In : *The Journal of the Acoustical Society of America* 123.5, p. 3163-3163. doi : 10.1121/1.2933214.

— (mai 2015). « The Development of a Sound Wheel for Reproduced Sound ». In : *Audio Engineering Society Convention 138*. url : http://www.aes.org/e-lib/browse.cfm?elib=17734.

Picinali, Lorenzo et al. (2014). « Exploration of architectural spaces by blind people using auditory virtual reality for the construction of spatial knowledge ». In : *International Journal of Human-Computer Studies* 72.4, p. 393-407.

Pick, Herbert L, David H Warren et John C Hay (1969). « Sensory conflict in judgments of spatial direction ». In : *Perception & Psychophysics* 6.4, p. 203-205.

Pinson, Margaret H et Stephen Wolf (2003). « Comparing subjective video quality testing methodologies ». In : *Visual Communications and Image Processing 2003*. T. 5150. International Society for Optics et Photonics, p. 573-583.

Plomp, R. (1976). *Aspects of tone sensation : a psychophysical study*. Academic press series in cognition and perception /ed. by Edward Carterette, Morton P. Friedman. Academic Press. isbn : 9780125583503. url : https://books.google.fr/books?id=DOpqAAAAMAAJ.

Pulkki, Ville et Toni Hirvonen (2005). « Localization of virtual sources in multichannel audio reproduction ». In : *IEEE Transactions on Speech and Audio Processing* 13.1, p. 105-119.

Raake, Alexander (2007). *Speech quality of VoIP : assessment and prediction*. John Wiley & Sons.

— (2016). « Views on Sound Quality ». In : *Proceedings of the 22th International Congress of Acoustics, Buenos Aires, Argentina*.

Raake, Alexander et Sebastian Egger (2014). « Quality and quality of experience ». In : *Quality of Experience*. Springer, p. 11-33.

Radeau, Monique et Paul Bertelson (1977). « Adaptation to auditory-visual discordance and ventriloquism in semirealistic situations ». In : *Perception & Psychophysics* 22.2, p. 137-146.

Reiter, Ulrich et al. (2014). « Factors influencing quality of experience ». In : *Quality of Experience*. Springer, p. 55-72.

Riggs, William et Kayla Gordon (2017). « How is mobile technology changing city planning ? Developing a taxonomy for the future ». In : *Environment and Planning B : Urban Analytics and City Science* 44.1, p. 100-119.

Robotham, Thomas et al. (2018). « Online vs. Offline Multiple Stimulus Audio Quality Evaluation for Virtual Reality ». In : *Audio Engineering Society Convention 145*. Audio Engineering Society.

Roto, Virpi et al. (2006). *Web browsing on mobile phones : Characteristics of user experience*. Helsinki University of Technology.

Rubino, Gerardo, Pierre Tirilly et Martın Varela (2006). « Evaluating users' satisfaction in packet networks using random neural networks ». In : *International Conference on Artificial Neural Networks*. Springer, p. 303-312.

Rugeles Ospina, F, Marc Emerit et Jérôme Daniel (2015). « A fast measurement of high spatial resolution head related transfer functions for the bili project ». In : *Proc. 3rd International Conference on Spatial Audio*.

Rummukainen, Olli et al. (2018). « Influence of visual content on the perceived audio quality in virtual reality ». In : *Audio Engineering Society Convention 145*. Audio Engineering Society.

Rumsey, Francis (2002). « Spatial Quality Evaluation for Reproduced Sound : Terminology, Meaning, and a Scene-Based Paradigm ». In : *Journal of the Audio Engineering Society* 50.9, p. 651-666. ISSN : 00047554.

Rumsey, Francis et al. (2005). « On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality. » In : *The Journal of the Acoustical Society of America* 118.2, p. 968-976. DOI : 10.1121/1.1945368.

Schilit, Bill N, Norman Adams et Roy Want (1994). « Context-aware computing applications ». In : *Proceedings of the 1994 First Workshop on Mobile Computing Systems and Applications*, p. 85-90. ISSN : 15277755. DOI : 10.1109/MCSA.1994.512740. URL : http://ieeexplore.ieee.org/xpls/abs%7B%5C_%7Dall.jsp?arnumber=4624429.

Schleicher, Robert, Tilo Westermann et Ralf Reichmuth (2014). « Mobile Human–Computer Interaction ». In : *Quality of Experience*. Springer, p. 339-349.

Schöffler, Michael (2017). « Overall listening experience—A new approach to subjective evaluation of audio ». Thèse de doct. International Audio Laboratories Erlangen.

Schönstein, David et Brian FG Katz (2012). « Variability in perceptual evaluation of HRTFs ». In : *Journal of the Audio Engineering Society* 60.10, p. 783-793.

Schuurman, Dimitri, Lieven De Marez et Pieter Ballon (2015). « Living Labs : a systematic literature review ». In : *Open Living Lab Days 2015*.

Schuurman, Dimitri, Katrien De Moor et al. (2011). « A Living Lab research approach for mobile TV ». In : *Telematics and Informatics* 28.4, p. 271-282.

Scriven, Frances (2005). « Two types of sensory panels or are there more ? » In : *Journal of sensory studies* 20.6, p. 526-538.

Seebode, Julia, Robert Schleicher et Sebastian Möller (2012). « Affective quality of audio feedback in different contexts ». In : *Proceedings of the 11th International Conference on Mobile and Ubiquitous Multimedia*. ACM, p. 32.

Sekuler, Robert, Allison B Sekuler et Renee Lau (1997). « Sound alters visual motion perception. » In : *Nature* 385.6614, p. 308.

Semi-Secret Software (2009). *Canabalt*. [iOS].

Shams, Ladan, Yukiyasu Kamitani et Shinsuke Shimojo (2000). « Illusions : What you see is what you hear ». In : *Nature* 408.6814, p. 788.

Shinn-Cunningham, Barbara G, Timothy Streeter et Jean-François Gyss (2005). « Perceptual plasticity in spatial auditory displays ». In : *ACM Transactions on Applied Perception (TAP)* 2.4, p. 418-425.

Shirley, Ben Guy et al. (2017). « Personalized object-based audio for hearing impaired TV viewers ». In : *Journal of the Audio Engineering Society* 65.4, p. 293-303.

Shively, Roger (1998). « Subjective evaluation of reproduced sound in automotive spaces ». In : *Audio Engineering Society Conference : 15th International Conference : Audio, Acoustics & Small Spaces*. Audio Engineering Society.

Simon, Laurent SR, Nick Zacharov et Brian FG Katz (2016). « Perceptual attributes for the comparison of head-related transfer functions ». In : *The Journal of the Acoustical Society of America* 140.5, p. 3623-3632.

Slutsky, Daniel A et Gregg H Recanzone (2001). « Temporal and spatial dependency of the ventriloquism effect ». In : *Neuroreport* 12.1, p. 7-10.

Smith, D, L Ma et N Ryan (2006). « Acoustic environment as an indicator of social and physical context ». In : *Personal and Ubiquitous Computing*. URL : http://link.springer.com/article/10.1007/s00779-005-0045-4.

Sohn, Timothy et al. (2008). « A diary study of mobile information needs ». In : *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, p. 433-442.

Somethin'Else (2011a). *Papa Sangre*. [iOS].

— (2011b). *The Nightjar*. [iOS].

— (2013). *Papa Sangre 2*. [iOS].

Soriano, Delphine (2016). « Outils d'évaluation du design de l'avatar dans l'engagement du joueur ». Thèse de doct. Conservatoire national des arts et métiers - CNAM.

Spagnol, Simone, Michele Geronazzo et Federico Avanzini (2013). « On the relation between pinna reflection patterns and head-related transfer function features ». In : *IEEE transactions on audio, speech, and language processing* 21.3, p. 508-519.

Spence, Charles et Jon Driver (2000). « Attracting attention to the illusory location of a sound : reflexive crossmodal orienting and ventriloquism ». In : *Neuroreport* 11.9, p. 2057-2061.

Spors, Sascha et al. (2013). « Spatial sound with loudspeakers and its perception : A review of the current state ». In : *Proceedings of the IEEE* 101.9, p. 1920-1938. ISSN : 00189219. DOI : 10.1109/JPROC.2013.2264784.

Staelens, Nicolas et al. (2014). « Subjective quality assessment of longer duration video sequences delivered over HTTP adaptive streaming to tablet devices ». In : *IEEE Transactions on Broadcasting* 60.4, p. 707-714.

Staffeldt, Henrik (1974). « Correlation between subjective and objective data for quality loudspeakers ». In : *Audio Engineering Society Convention 47*. Audio Engineering Society.

Stein, Barry E et Terrence R Stanford (2008). « Multisensory integration : current issues from the perspective of the single neuron ». In : *Nature Reviews Neuroscience* 9.4, p. 255-266.

Stone, H et al. (1974). « Sensory evaluation by quantitative descriptive analysis ». In : *Food technology*.

Stratton, George M (1897). « Vision without inversion of the retinal image. » In : *Psychological review* 4.4, p. 341.

Strohmeier, Dominik, Satu Jumisko-Pyykkö et Kristina Kunze (2010). « Open profiling of quality : a mixed method approach to understanding multimodal quality perception ». In : *Advances in multimedia* 2010.

Talsma, Durk et al. (2010). « The multifaceted interplay between attention and multisensory integration ». In : *Trends in cognitive sciences* 14.9, p. 400-410.

Tamminen, Sakari et al. (2004). « Understanding mobile contexts ». In : *Personal and ubiquitous computing* 8.2, p. 135-143.

Teunissen, Kees et Joyce HDM Westerink (1996). « A multidimensional evaluation of the perceptual quality of television sets ». In : *SMPTE journal* 105.1, p. 31-38.

Thurlow, Willard R et Charles E Jack (1973). « Certain determinants of the "ventriloquism effect" ». In : *Perceptual and motor skills* 36.3_suppl, p. 1171-1184.

Toole, Floyd E (1985). « Subjective measurements of loudspeaker sound quality and listener performance ». In : *Journal of the Audio Engineering Society* 33.1/2, p. 2-32.

Torcoli, Matteo et al. (2017). « The adjustment/satisfaction test (A/ST) for the subjective evaluation of dialogue enhancement ». In : *Audio Engineering Society Convention 143*. Audio Engineering Society.

Traer, James et Josh H McDermott (2016). « Statistics of natural reverberation enable perceptual separation of sound and space ». In : *Proceedings of the National Academy of Sciences* 113.48, E7856-E7865.

Tucker, I (2011). « Perceptual video quality dimensions ». Thèse de doct. Master thesis, Technische Universität Berlin, Berlin.

Valve Corporation (1998). *Half-Life*. PC (Windows).

Van der Burg, Erik, John Cass et al. (2010). « Efficient visual search from synchronized auditory signals requires transient audiovisual events ». In : *PLoS One* 5.5, e10664.

Van der Burg, Erik, Christian NL Olivers et al. (2008). « Pip and pop : nonspatial auditory signals improve spatial visual search. » In : *Journal of Experimental Psychology : Human Perception and Performance* 34.5, p. 1053.

Varela, Martín, Lea Skorin-Kapov et Touradj Ebrahimi (2014). « Quality of service versus quality of experience ». In : *Quality of Experience*. Springer, p. 85-96.

Vroomen, Jean, Paul Bertelson et Beatrice De Gelder (2001). « The ventriloquist effect does not depend on the direction of automatic visual attention ». In : *Perception & psychophysics* 63.4, p. 651-659.

Vurpillot, Eliane (1963). « La perception de l'espace ». In : *Traité de Psychologie Expérimentale. Piaget J y cols.(eds.). PUF. Francia.*

Wagner, Daniel T, Andrew Rice et Alastair R Beresford (2013). « Device analyzer : Understanding smartphone usage ». In : *International Conference on Mobile and Ubiquitous Systems : Computing, Networking, and Services*. Springer, p. 195-208.

Wallace, Mark T et al. (2004). « Unifying multisensory signals across time and space ». In : *Experimental Brain Research* 158.2, p. 252-258.

Wältermann, Marcel (2013). *Dimension-based quality modeling of transmitted speech*. Springer Science & Business Media.

Walton, Tim et Michael Evans (2018). « The role of human influence factors on overall listening experience ». In : *Quality and User Experience* 3.1, p. 1.

Walton, Tim, Michael Evans et al. (2018). « Exploring object-based content adaptation for mobile audio ». In : *Personal and Ubiquitous Computing* 22.4, p. 707-720.

Walton, Timothy (2018). « The quality of experience of next generation audio : exploring system, context and human influence factors ». Thèse de doct. Newcastle University.

Wang, Shaoxuan et Sujit Dey (2012). « Cloud mobile gaming : Modeling and measuring user experience in mobile wireless networks ». In : *ACM SIGMOBILE Mobile Computing and Communications Review* 16.1, p. 10-21.

Warren, David H, Robert B Welch et Timothy J McCarthy (1981). « The role of visual-auditory "compellingness" in the ventriloquism effect : Implications for transitivity among the spatial senses ». In : *Perception & Psychophysics* 30.6, p. 557-564.

Warusfel, O (2003). *Listen HRTF Database*. recherche.ircam.fr/equipes/salles/listen.

Wechsung, Ina et Katrien De Moor (2014). « Quality of Experience Versus User Experience ». In : *Quality of Experience*. Springer, p. 35-54.

Weiss, Benjamin et al. (2014). « Temporal development of quality of experience ». In : *Quality of experience*. Springer, p. 133-147.

Welch, Robert B et David H Warren (1980). « Immediate perceptual response to intersensory discrepancy. » In : *Psychological bulletin* 88.3, p. 638.

Wenzel, Elizabeth M et al. (1993). « Localization using nonindividualized head-related transfer functions ». In : *The Journal of the Acoustical Society of America* 94.1, p. 111-123.

Werner, Stephan et Florian Klein (2014). « Influence of context dependent quality parameters on the perception of externalization and direction of an auditory event ». In : *Audio Engineering Society Conference : 55th International Conference : Spatial Audio*. Audio Engineering Society.

Werner, Stephan, Judith Liebetrau et Thomas Sporer (2013). « Vertical sound source localization influenced by visual stimuli ». In : *Signal Processing Research* 2.2, p. 29-38.

Wigelius, Heli et Heli Väätäjä (2009). « Dimensions of context affecting user experience in mobile work ». In : *IFIP Conference on Human-Computer Interaction*. Springer, p. 604-617.

Wightman, Frederic L et Doris J Kistler (1989). « Headphone simulation of free-field listening. II : Psychophysical validation ». In : *The Journal of the Acoustical Society of America* 85.2, p. 868-878.

— (1992). « The dominant role of low-frequency interaural time differences in sound localization ». In : *The Journal of the Acoustical Society of America* 91.3, p. 1648-1661.

— (1999). « Resolution of front–back ambiguity in spatial hearing by listener and source movement ». In : *The Journal of the Acoustical Society of America* 105.5, p. 2841-2853.

Willey, Clarence F, Edward Inglis et CH Pearce (1937). « Reversal of auditory localization. » In : *Journal of Experimental Psychology* 20.2, p. 114.

Winkler, Stefan et Frédéric Dufaux (2003). « Video quality evaluation for mobile streaming applications ». In : *Visual Communications and Image Processing 2003*. T. 5150. International Society for Optics et Photonics, p. 593-604.

Winkler, Stefan et Christof Faller (2006). « Perceived audiovisual quality of low-bitrate multimedia content ». In : *IEEE transactions on multimedia* 8.5, p. 973-980.

Witkin, Herman A, Seymour Wapner et Tama Leventhal (1952). « Sound localization with conflicting visual and auditory cues. » In : *Journal of experimental psychology* 43.1, p. 58.

Witmer, Bob G et Michael J Singer (1998). « Measuring presence in virtual environments : A presence questionnaire ». In : *Presence* 7.3, p. 225-240.

Woodworth, Robert Sessions et Harold Schlosberg (1954). *Experimental psychology*. Oxford et IBH Publishing.

Wright, Richard D et Lawrence M Ward (2008). *Orienting of attention*. Oxford University Press.

Wu, Wanmin et al. (2009). « Quality of experience in distributed interactive multimedia environments : toward a theoretical framework ». In : *Proceedings of the 17th ACM international conference on Multimedia*. ACM, p. 481-490.

Yamagishi, Kazuhisa et Takanori Hayashi (2005). « Analysis of psychological factors for quality assessment of interactive multimodal service ». In : *Human Vision and Electronic Imaging X*. T. 5666. International Society for Optics et Photonics, p. 130-139.

Yates, Frances Amelia et Daniel Arasse (1987). *L'art de la mémoire*. Gallimard.

Young, Paul Thomas (1928). « Auditory localization with acoustical transposition of the ears. » In : *Journal of Experimental Psychology* 11.6, p. 399.

Zacharov, Nick et Kalle Koivuniemi (2001a). « Unravelling the perception of spatial sound reproduction : Analysis & external preference mapping ». In : *Audio Engineering Society Convention 111*. Audio Engineering Society.

— (2001b). « Unravelling the perception of spatial sound reproduction : Language development, verbal protocol analysis and listener training ». In : *Audio Engineering Society Convention 111*. Audio Engineering Society.

— (2001c). « Unravelling the perception of spatial sound reproduction : Techniques and experimental design ». In : *Audio Engineering Society Conference : 19th International Conference : Surround Sound-Techniques, Technology, and Perception*. Audio Engineering Society.

Zacharov, Nick, Torben Holm Pedersen et Chris Pike (2016). « A common lexicon for spatial sound quality assessment-latest developments ». In : *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, p. 1-6.

Zadtootaghaj, Saman, Steven Schmidt et Sebastian Möller (2018). « Modeling Gaming QoE : Towards the Impact of Frame Rate and Bit Rate on Cloud Gaming ». In : *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, p. 1-6.

Zahorik, Pavel (2003). « Auditory and visual distance perception : The proximity-image effect revisited ». In : *The Journal of the Acoustical Society of America* 113.4, p. 2270-2270.

Zahorik, P et al. (2001). « Localization accuracy in 3D sound displays : The role of visual-feedback training ». In : *Proceedings of the advanced displays and interactive displays federal laboratory consortium*.

Zhang, Peter Xinya et William M Hartmann (2010). « On the ability of human listeners to distinguish between front and back ». In : *Hearing research* 260.1-2, p. 30-46.

Zheng, Yong et Alisha Anna Jose (2019). « Context-Aware Recommendations via Sequential Predictions ». In :

Zielinski, Slawomir, Peter Brooks et Francis Rumsey (oct. 2007). « On the Use of Graphic Scales in Modern Listening Tests ». In : *Audio Engineering Society Convention 123*. URL : http://www.aes.org/e-lib/browse.cfm?elib=14234.

Julian Moreira

Évaluer l'apport du binaural dans une application mobile audiovisuelle

Résumé

Dans ce travail de thèse, nous nous proposons d'évaluer l'intérêt que peut présenter le son binaural lorsqu'il est utilisé dans une application mobile audiovisuelle. Dans une première partie, nous nous intéressons à la perception spatiale d'un son binaural associé à un visuel rendu sur un petit écran. Notre expérience révèle une forte tolérance de l'être humain face aux dégradations spatiales pouvant survenir entre les deux modalités. Dans la seconde partie, nous répondons à la question de l'apport du binaural en termes d'immersion, de mémorisation et de performance, dans un jeu vidéo mobile de type *Infinite Runner*. Nous déployons une expérience « hors les murs », dans un contexte plausible d'utilisation grand public. Les résultats indiquent une immersion significativement meilleure pour le binaural, comparée à un rendu monophonique. La mémorisation et la performance ne sont en revanche pas soumises à un effet statistiquement significatif du rendu sonore. Au-delà des résultats, cette expérience nous permet de discuter de la question de la validité des données en fonction de la méthode de déploiement, en confrontant notamment bienfondé théorique et faisabilité pratique.

Mots clés : Binaural, smartphone, qualité d'expérience, expérience utilisateur, perception audiovisuelle, effet ventriloque, point d'alignement spatial subjectif, méthode d'échantillonnage de l'expérience, contexte, attribut sonore, jeu vidéo

Abstract

In this thesis, we study the potential contribution of the binaural technology to enhance the quality of experience of an audiovisual mobile application. In the first part, we address the question of the spatial perception of a binaural sound associated to a visual rendered on a small screen. The results of our experiment reveal a strong tolerance of subjects to spatial discrepancies between the two modalities. In the second part, we address the contribution of the binaural rendering in terms of immersion, memorization and performance, in an *Infinite Runner*, a widespread smartphone video game type. In particular, we conduct the experiment in a realistic context of use. Results indicate a significantly better immersion with the binaural rendering, compared to the monophonic one. No effect of sound rendering was found for memorization and performance. Beyond the contribution of the binaural, we discuss about the protocol, the validity of the collected data, and oppose theoretical considerations to practical feasibility.

Keywords : Binaural sound, smartphone, quality of experience, user experience, audiovisual perception, ventriloquist effect, point of subjective spatial alignment, experience sampling method, context, sound attribute, video game