Louis Lumière
école nationale supérieure

Un auditorium cinéma en binaural : transfert du mixage de la salle au casque

Jeanne GRIVELET

Mémoire de Master 2 – Spécialité Son

Directeur de mémoire interne : Sylvain LAMBINET

Directeur de mémoire externe : Etienne HENDRICKX

Responsable universitaire : Corsin VOGEL

Juin 2021

Résumé

Le marché des plug-ins et certains mixeurs se sont emparés de la technologie binaurale dans le but de recréer des environnements de travail conséquents et réalistes. Cela permettrait de mixer au casque sans subir son gros défaut qu'est la stéréophonie. En effet, notre écoute en champ libre est bien différente de celle dont nous avons l'habitude au casque, peu adaptée pour une écoute de son surround. Si cette nouvelle technique de mixage s'avérait pertinente tant d'un point de vue du workflow que des bandes-son produites, elle pourrait s'implémenter durablement dans le paysage audiovisuel. La possibilité de mixer au casque dans un lieu autre que l'auditorium pourrait présenter de nombreux avantages dans des étapes comme le montage son ou la vérification de mix.

Ce mémoire vise à répondre à des questions que se posent les professionnels du son à l'image : est-il possible de mixer un film dans des conditions d'auditorium depuis chez soi ? Les résultats sont-ils convaincants ? Afin de répondre à ces problématiques, un protocole sera mis en place pour créer un auditorium en binaural et permettre à des mixeurs de travailler dans cette condition, puis d'évaluer les résultats en les comparant avec des mixages effectués dans une condition in situ.

Abstract

The plug-in market and some mixers have seized on binaural technology in order to recreate consistent and realistic playback environments. This would make it possible to mix with headphones without being reduced to stereophony which doesn't fit with surround sound reproduction. Indeed, our free field listening experience is quite different from what we are used to with headphones. If this new mixing technique proves to be relevant both from a workflow point of view and from the mixing quality, it could be implemented permanently in the audiovisual landscape. The possibility of mixing with headphones in a place other than the auditorium could have many advantages in stages such as sound editing or mix checking.

This thesis aims at answering questions that sound and image professionals ask themselves: is it possible to mix a film in auditorium conditions from home? Are the results convincing? In order to answer these questions, a protocol will be set up to create a binaural auditorium and allow mixers to work in this condition, then to evaluate the results by comparing them with mixes made in situ.

Mots-clés : post-production, mixage, casque, binaural, perception

Remerciements

Je souhaite remercier Cyril Holtz pour sa proposition de sujet, sans laquelle je ne serais sûrement pas en train d'écrire ces lignes.

Merci à Etienne Hendrickx pour m'avoir guidée dans les domaines merveilleux du binaural, des tests perceptifs et des statistiques.

Merci infiniment à François Salmon et Charles Verron pour avoir consacré du temps à la confection des HRTF et qui ont donc permis toute la partie expérimentale de ce mémoire.

Un grand merci à Boris Chappelle pour avoir pris le temps de me parler de son expérience et de son avis sur la question.

Merci beaucoup à Alan Blum pour m'avoir dégoté une formidable KU-100, qui m'a gracieusement été prêtée par l'équipe de Brian Katz à Jussieu, merci à eux également.

Merci énormément à Aloyse Launay et Dimitri Kharitonnoff pour leur patience et leur dévouement dans cet ultime projet étudiant bénévole.

Merci à Martin Guesney pour son support inconditionnel, et aussi (surtout) pour sa 308.

Merci à toutes les personnes qui ont subi les tests perceptifs, la Science s'en trouve grandie grâce à vous.

Merci à Sylvain Lambinet et Corsin Vogel pour leurs précieux conseils et leur accompagnement durant cette longue et laborieuse période.

Merci à mes parents.

Table des matières

Introduction

Il n'a échappé à personne que ces derniers temps ont poussé les travailleurs à se retrancher chez eux et à établir le moins de contacts physiques possibles avec d'autres personnes. Les studios étant des lieux de travail consacrés au son, mais également des lieux de rencontres et d'échanges, ne sont pas idéaux dans un contexte de pandémie. Les salles de cinéma ont également fermé leurs portes pour plusieurs mois, ce qui a permis aux services de vidéos à la demande de faire fructifier leurs entreprises et habituer encore plus le consommateur à regarder et écouter tout contenu médiatique depuis son domicile.

Heureusement pour cette situation, les technologies ont toujours suivi cette tendance inexorable à la miniaturisation et l'individualisation, qui a produit des systèmes toujours plus mobiles, petits et personnels. Ces systèmes tranchent avec ceux traditionnellement associés au cinéma et au son, que sont les auditoriums pourvus d'un écran de plusieurs mètres de diagonale et d'une console tout aussi large, dans une salle qui doit pouvoir les contenir en permettant une bonne distance de recul.

Le mixeur Cyril Holtz a proposé cette problématique qu'il creuse avec son studio depuis quelques temps, qui est celle du transfert de mixage d'une salle (qu'elle soit de montage ou autre) vers l'auditorium. En effet, la tendance veut que de plus en plus d'étapes auparavant dédiées au mixage soient reléguées à l'étape du montage son, donc dans un endroit qui ne soit pas aussi équipé et traité qu'un auditorium. Le prémixage est une étape du montage son qui consiste à préparer la bande-son au mixage, en effectuant un mixage basique (ou poussé, selon les moyens et besoins). Le prémixage gagnerait donc à être également pré-écouté dans l'auditorium qui le recevra.

Une des solutions les plus simples est donc d'avoir recours au binaural, technologie connue de longue date pour sa simplicité et son efficacité relativement au dispositif requis pour plonger l'auditeur dans un environnement multidimensionnel convaincant. Il y a donc une demande pour des dispositifs au casque, qui voit offre chez plusieurs constructeurs hardware ou software. Un tel dispositif pourrait également être utilisé par des réalisateurs dans un cadre d'écoute à distance, pour qu'ils retrouvent les sensations de ce qu'ils avaient entendu en salle, ou simplement avoir une meilleure idée de l'avancée d'un montage ou mixage son tel qu'il a été écouté par le monteur ou le mixeur.

Cependant, ce n'est pas parce qu'une technologie se propose qu'elle est efficace, ou que sa visée soit louable. Il convient donc de déterminer si un travail de mixage au casque est envisageable ou souhaitable, et que les résultats de ces mixages sont d'une qualité et d'une transparence équivalentes à leurs homologues d'auditorium. Nous allons donc observer des mixeurs travailler dans ces conditions dont ils n'ont pas forcément l'habitude, puis comparer pour un même extrait de film les résultats des deux conditions

qui nous intéressent. Comme l'écoute est subjective, nous procéderons à des tests perceptifs sur un panel d'experts pour déterminer la qualité des mixages.

Les problématiques qui vont motiver ce mémoire sont donc les suivantes : est-ce possible de recréer un auditorium en binaural ? Comment se déroule l'insertion dans le workflow ? Le mixage qui résulte d'un tel dispositif binaural est-il à la hauteur de ce qu'on peut en attendre pour son équivalent in situ ?

Dans un premier temps, nous rappellerons les fonctions et particularités de l'écoute. L'étude de la physiologie nous permettra ensuite d'aborder la technologie binaurale, ainsi que les différentes solutions existantes dans ce domaine. Nous établirons alors un protocole de test perceptif en vue d'évaluer la robustesse d'un mixage entre sa version in situ et sa version binaurale, que nous conclurons sur ses résultats.

Partie 1 : Etat de l'art

1 – Perception du son spatial

L'être humain perçoit son environnement sonore en trois dimensions. Cela est permis par la nature de la propagation des ondes sonores qui sont réfléchies par diverses surfaces dans tous les axes possibles autour de l'auditeur. La recherche de cette sensation naturelle de l'immersion a toujours été au cœur de l'avancée des techniques sonores afin de proposer une expérience au plus près du réel au sein de différents média. Pour recréer au mieux notre expérience du son en milieu naturel, il convient d'appréhender dans un premier temps la manière dont nous percevons les sons par tout son appareil physiologique et cognitif.

1 – 1. Physiologie de l'oreille

La perception auditive est un phénomène complexe qui repose sur la physiologie du système auditif et le traitement de ces données reçues par le cerveau dans des processus cognitifs. Avant de décortiquer les informations collectées par l'oreille en des paramètres compréhensibles tels que le timbre, la spatialisation ou la dynamique, il convient de comprendre comment ces données sont formées en arrivant au cerveau. Ces deux fonctions sont appelées processus périphérique et processus central [RG17].

Le système auditif périphérique comprend toutes les étapes de l'oreille externe au nerf cochléaire. L'oreille est composée de trois parties distinctes que sont l'oreille externe, l'oreille moyenne et l'oreille interne.

L'oreille externe comprend le pavillon et le conduit auditif externe. La forme de cornet du pavillon permet de concentrer les ondes sonores vers l'oreille moyenne située au bout du conduit auditif. La longueur du conduit auditif fait qu'il y a une résonance pour les fréquences autour de 3kHz (pression 30 à 100 fois plus importante), soit une acuité plus élevée pour les sons entre 2kHz et 5kHz, ce qui correspond au spectre d'intelligibilité de la voix. La forme du pavillon permet également de donner des informations sur la perception de l'élévation ou de la profondeur d'une source [S74]. La forme asymétrique du pavillon lui permet de transmettre plus de hautes fréquences lorsque la source se situe au-dessus de l'auditeur que si elle était au niveau de son oreille, de même que ces hautes fréquences sont atténuées lorsqu'elles se situent à l'arrière [B97].

L'oreille moyenne est une cavité intermédiaire composée de trois petits os que sont le marteau, l'enclume et l'étrier qui permettent d'adapter la résistance de l'air à la résistance du fluide de l'oreille interne, ce qui empêche une perte de 30dB dans la transmission. C'est le tympan, une membrane reliée à ces osselets qui vibre selon les ondes reçues et ainsi transmet le son à amplifier. Lorsque le niveau sonore dépasse les

85dB, un autre muscle appelé le muscle stapédien se contracte et ainsi rigidifie l'ensemble de la chaine des osselets pour atténuer la transmission sonore ; c'est le réflexe stapédien.

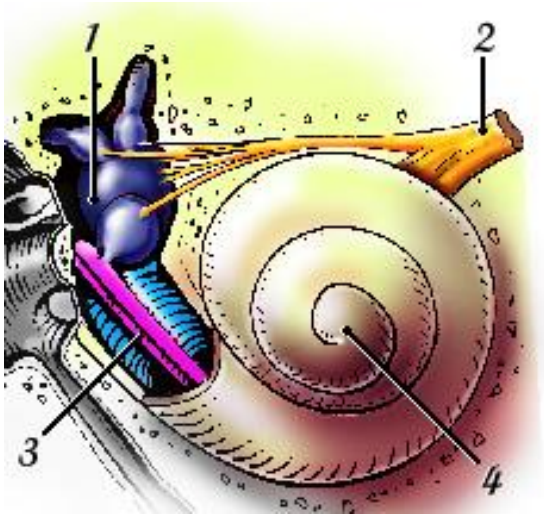


Figure 1 : Schéma de l'oreille interne, avec en (1) les organes vestibulaires, en (2) les nerfs vestibulaires et cochléaires, en (3) le tour basal de la cochlée et en (4) l'apex. Source : www.cochlea.eu

L'oreille interne a deux fonctions importantes : la gestion de l'équilibre et la transduction des signaux acoustiques en impulsions électriques. Cette deuxième fonction essentielle à l'ouïe est effectuée par la cochlée, organe creux rempli de liquides appelés endolymphe et périlymphe et enroulé sur lui-même. Deux membranes séparent la cochlée en trois compartiments : la membrane basilaire et la membrane de Reissher (ou membrane vestibulaire). L'organe de Corti, qui est recouvert de cellules ciliées permettant la transduction, se situe sur la membrane basilaire et est en contact avec la membrane tectoriale qui joue un rôle important dans la liaison et la bonne transmission des ondes sonores, notamment vérifie que la cochlée reçoive bien les ondes à la bonne intensité et au bon moment [RLR08]. Les cellules ciliées réagissent au mouvement des ondes et décomposent le signal selon sa fréquence, son amplitude et sa phase vers les fibres des nerfs auditifs. Cependant, la membrane basilaire n'étant pas uniforme dans sa rigidité, les fréquences basses sont reconnues vers le centre de la cochlée, alors que les hautes fréquences sont reconnues vers la base. Ainsi, la membrane basilaire peut être interprétée comme séparée en plusieurs types de filtres [PNHR87] [ML10]. Les fréquences seraient alors interprétées par les cellules selon leur place dans la cochlée [B60] [J48].

1 – 2. Localisation des sources sur le plan horizontal

La perception spatiale auditive renvoie à la capacité de localiser des sources dans un environnement 3D même lorsqu'elles sont multiples. Contrairement à la vue, les informations spatiales ne sont pas comprises directement par le système auditif mais sont déduites des propriétés binaurales et du filtrage fréquentiel opéré par le pavillon [YD97]. En trois dimensions, on place alors la source en termes d'azimut, d'élévation ou de

distance, soit pour un auditeur regardant face à lui l'azimut est à 0° et l'élévation à 0° également. Le plan azimut est donc le plan horizontal et le plan d'élévation le plan vertical, les positions à la droite et en haut de l'auditeur étant positives (gauche/bas sont négatives). La distance est définie selon l'azimut et l'élévation.

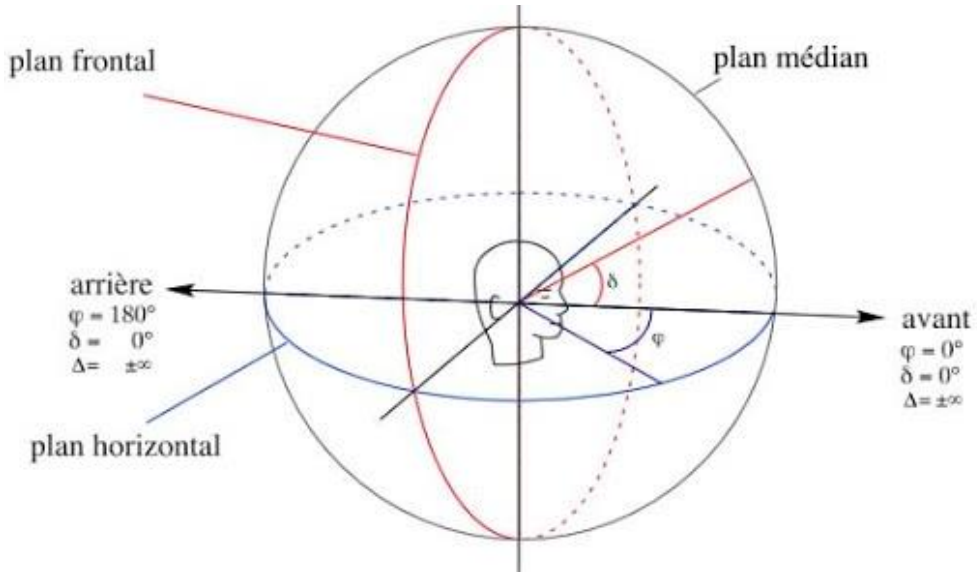


Figure 2 : Les différents plans de perception sonore. Source : unige.ch

Ce qui permettra à l'auditeur de définir la position azimutale d'un son est la différence en temps et en niveau perçue par les oreilles : on parle d'ITD (*interaural time differences*) et d'ILD (*interaural level differences*, aussi appelées IID pour *interaural intensity differences*). Pour définir son élévation et sa position avant/arrière, le système auditif se base sur le filtrage apporté par la forme du pavillon [B97] [HV03]. Le délai en temps dépend de la distance entre les deux oreilles, du volume crânien et de sa forme qui provoque des réflexions (qui jouent également sur la perception du niveau). La différence d'intensité s'explique par une source perçue plus fort si elle se trouve plus proche d'une oreille que de l'autre. Récemment, il a été démontré que les auditeurs sont sensibles aux ITD jusqu'à 1400Hz [BDH13]. Il est possible de simplifier la tête de l'auditeur par une sphère et deux oreilles pour représenter le trajet d'une onde qui se mue en deux chemins, chacun allant du point central de la source vers une oreille.

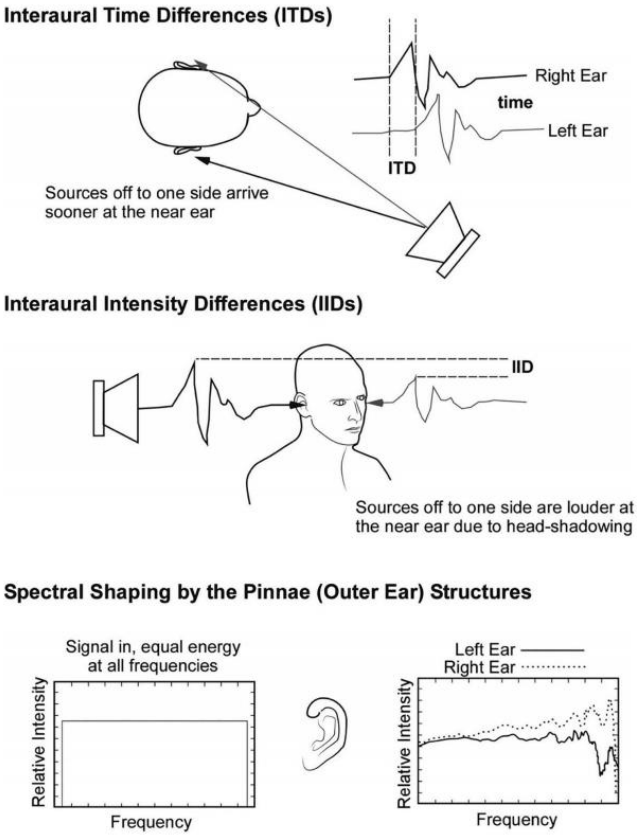


Figure 3 : Illustration des différents indices de localisation (crédit : Roginska & Geluso, Immersive Sound, 2018)

A 90° azimut, le maximum d'ITD perçu est de 700µs environ. Ces mesures varient de façon linéaire entre 0° et 90° puis entre 90° et 180° de 0 à 700µs. Le système auditif comparerait les pics dans les excitations neuronales entre les données venant de l'oreille gauche et celles venant de l'oreille droite pour définir les ITD, les neurones attribués aux ITD d'un hémisphère étant corrélés à ceux de l'autre [B62] [MK10] [STYM10].

Là où les ITD sont essentielles pour localiser des fréquences inférieures à 2000Hz, les ILD jouent un rôle plus important lorsque les longueurs d'onde sont plus petites que le diamètre du crâne (au-dessous elles sont diffractées donc moins atténuées). La tête agit alors comme un obstacle qui empêche les ondes de passer à l'autre oreille. Plus la fréquence sera élevée et donc la longueur d'onde petite, plus l'effet sera important. La perte d'intensité peut aller jusqu'à plus de 20dB pour 10kHz à un angle azimutal de 90°. Cette perte d'intensité fonctionne de manière indépendante du changement fréquentiel pour déterminer l'emplacement d'une source (voir le potentiomètre de pan qui fonctionne uniquement sur une distribution de niveau) car est efficace sur tout le spectre jusqu'à au moins 200Hz.

Les recherches sur le binaural tendent cependant à montrer que cette division de différenciation fréquentielle ou *duplex theory*, soit ILD pour les hautes fréquences, ITD pour les basses, n'est pas si nette que nous pouvons le croire. L'ILD serait valable pour

les basses fréquences à de courtes distances, de même que pour les hautes fréquences les ITD seraient basées sur le délai relatif des amplitudes des enveloppes sonores. Il serait également possible de déterminer une source en mouvement en azimut à l'aide d'une seule oreille par la simple propriété des hautes fréquences à être plus atténuées que les basses [SDC08].

1 – 3. Localisation des sources sur le plan vertical

La distance interaurale joue peu dans la localisation verticale des sources (si les sources sont placées à un degré d'azimut nul), ce qui mène à une latéralisation ressentie à l'intérieur du crâne dans le cas où l'auditeur utilise un casque, où les ITD et ILD servent à percevoir la position des sources. L'élévation serait donc déterminée grâce à la forme du pavillon et de la réflexion des ondes sur le torse et les épaules. De légères asymétries crâniennes permettraient également de donner de faibles indices sur l'élévation. Le phénomène d'atténuation des hautes fréquences a lieu également dans la perception verticale des sons, car pour une même position un groupe de hautes fréquences centrées autour de 8kHz sera plus atténué qu'un groupe centré autour de 6kHz. L'absence de pavillon entrave la précision de localisation des sources et l'impression d'externalisation des sources [GG73] [OP84] [P74] [S74]. Une oreille seule peut également déterminer la quantité de réverbération apportée dans un signal, c'est-à-dire reconnaître un signal direct des ondes réfléchies par l'environnement pour une même source.

1 – 4. Problèmes liés à la localisation

Notre capacité à localiser les sons n'est cependant pas parfaite. Lors d'études qui demandaient à des auditeurs de situer des sources fixes en champ libre, ces derniers donnaient des emplacements de source qui pouvaient dévier entre 5° et 20° de la position originale : c'est le flou de localisation [B97]. Cette notion dépend également de l'azimut d'où elle provient : l'angle minimum (MAA, *Minimum Audible Angle*) pour détecter deux sources successives dans l'axe est de 1°, et de 20° pour des sources situées aux extrêmes latéraux [M58] [PS90]. Pour une source en mouvement continu, cet angle minimum perçu (MAMA, *Minimum Audible Movement Angle*) dépend de la vitesse de cette source et peut aller de 8° pour une vitesse de 90°/s à 21° pour une vitesse de 360°/s.

Le cône de confusion [W38] [WS54] [M72] amène l'auditeur à confondre des positions et à les inverser selon un axe avant/arrière ou haut/bas. Cela peut être expliqué par des ITD et ILD (quasi) exactement identiques pour deux sources, par exemple une source à 30° et une autre à 150° sur un plan azimutal à égale distance de l'oreille gauche, qui produisent alors le même effet de localisation.

Des informations supplémentaires sont alors nécessaires pour dissiper cet effet de cône de confusion, et peuvent être obtenues par la rotation et le déplacement de la tête [W39] [TR67] [WK99] [BWA01] ce qui modifie les rapports d'ITD et d'ILD. Le contenu spectral d'un son permet aussi de mieux le localiser : un son pleine bande sera plus facile

à localiser qu'un signal pur. La familiarité de l'auditeur avec certains types de son permet également de les localiser plus facilement, à force d'habitude : la réponse en fréquence d'une voix est généralement plus vite et mieux reconnue en distance et élévation [B97] [C63]. Cependant les meilleurs indices spectraux qui permettent de déterminer la position du son pour de mêmes ITD et ILD donnés restent les filtrages effectués par les oreilles qui, dû à la forme du pavillon, ne sont pas identiques pour des sources venant de l'avant ou l'arrière ou le haut et le bas.

1 – 5. Les HRTF

Les études ont montré qu'il est possible de recréer de manière relativement réaliste la localisation des sources si le filtrage effectué par l'oreille externe et les parties du corps influant sur la perception sonore est reproduit dans un système virtuel ou 3D. Ce filtre peut être mesuré par une impulsion enregistrée au niveau des oreilles par des microphones placés dans les conduits auditifs. Si les prises sont faites au même moment par les deux oreilles, on a alors les données d'ITD également. Cette technique permet de mesurer tous les indices spatiaux pour un emplacement de source donné, un auditeur donné et une salle ou environnement donné.

Ces filtres sont appelés HRIR, ou *Head-Related Impulse Responses* en temps et HRTF ou *Head-Related Transfer Functions* en fréquence, où les HRTF sont des transformées de Fourier des HRIR. Elles sont considérées comme des filtres RIF ou à Réponse Impulsionnelle Finie. Ces filtres comprennent donc la partie déterminée par la morphologie de l'auditeur, en particulier la forme des éléments externes avant le conduit auditif [M92], appelée DDF ou *Directional Dependant Function*, ainsi que les ILD et ITD. Ces HRTF sont donc propres à chaque individu, qui apprend à les connaître au fil de son existence et de ses expériences. La capacité de localisation évolue donc avec le temps et les circonstances.

1 – 6. Processus cognitifs

La perception auditive n'est pas qu'une affaire de signaux externes et d'appareil auditif mais comporte aussi toute une chaîne de traitement neuronale.

Le cerveau est sujet à la plasticité cérébrale, et toute la partie attribuée à l'audition ne fait pas exception : la capacité à localiser les sons peut évoluer en fonction des stimuli qu'elle reçoit et de son environnement. L'adaptation à de nouvelles données sensorielles peut même se faire rapidement. Il est même possible d'adapter un auditeur à des ILD et ITD correspondantes à l'obstruction d'une oreille, qui alors compense le manque au niveau du cortex auditif. L'auditeur fera alors plus confiance aux données inchangées et restées intactes en se basant sur l'oreille restée normale.

La plasticité permet cependant d'apprendre et de s'adapter à des HRTF différentes des nôtres, ce qui s'avèrera utile dans une implémentation à grande échelle de la technologie binaurale.

1 – 7. Psychoacoustique

1 – 7. 1. Quelques notions

La psychoacoustique est le point de rencontre des données physiques reçues par le corps et leur interprétation psychologique. Il est impossible de limiter l'écoute à un simple modèle de transducteur, puisque comme nous l'avons vu, le cerveau interprète en permanence les données.

La perception subjective de l'intensité acoustique est appelée sonie. Les caractéristiques de l'oreille amènent une perception différente de l'intensité acoustique par plages de fréquences, ce qui a permis de déduire des courbes d'isosonie pour un niveau de pression et une fréquence donnée. Ce sont des courbes d'égales sensations, en phones (un phone étant un niveau en dB SPL pour 1000Hz).

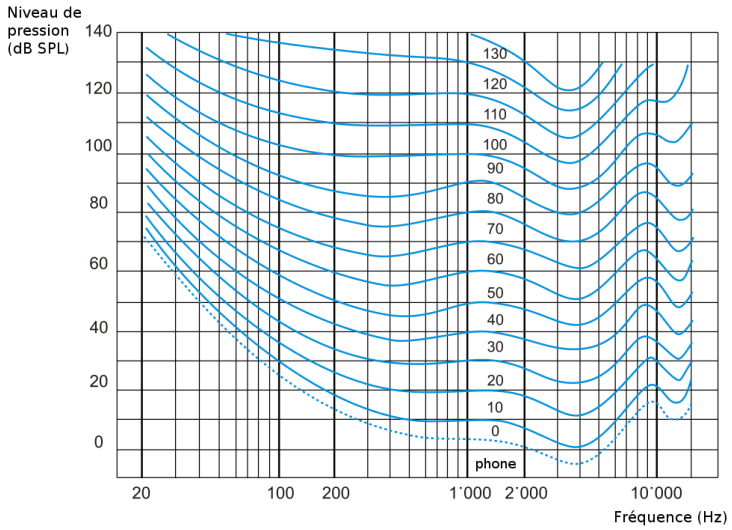


Figure 4 : Courbes d'isosonie selon la norme ISO 226 :2003. Source : wikipedia.org

Un son complexe n'aura pas ses composantes perçues de la même manière : on parle d'effet de masque. Les basses fréquences auront tendance à être perçues plus fortes et donc masquer les hautes fréquences, et cet effet est d'autant amplifié que l'intensité est élevée. Cet effet de masque fonctionne également en binaural, lorsqu'un bruit est présenté à une oreille et à l'autre un son pur : la sonie du son pur perçu sera modifiée.

L'effet de masque permet de dire que la perception des fréquences est également soumise à un jugement subjectif. La tonie n'est pas linéaire à partir de 500Hz, c'est-à-dire qu'il est plus difficile de faire la différence entre deux fréquences proches quand la fréquence augmente. La tonie varie également avec l'intensité : lorsqu'elle est élevée, les extrêmes sont exacerbés (graves perçus plus graves et aigus plus aigus). Telle la sonie,

c'est un phénomène binaural, qui se traduit par une estimation tonale différenciée entre les deux oreilles pour un même son pur : c'est la diplacousie.

1 – 7. 2. Liens entre vision et audition

Michel Chion [C02] parle d'aimantation spatiale pour désigner la localisation mentale, ou l'ajustement de notre capacité de localisation sonore à ce qui est vu à l'écran. La question la plus importante n'est pas de localiser précisément le son, mais plutôt celle de sa provenance (« d'où vient-il »). En effet, dans le cas d'une scène de marche, le spectateur aura l'impression que le son des pas suit les pas du personnage visible à l'écran, et semblent hors de l'écran s'il est hors-champ. L'arrivée du multicanal a amené un compromis entre la spatialisation mentale et la spatialisation « réelle », pourtant loin d'un réalisme le plus pur.

La zone de présence est un terme qui désigne la région où l'oreille localise et sépare le mieux lorsqu'elle s'associe à la vue [C95]. Elle permet de discriminer le son direct et le son réverbéré, ainsi qu'entre la partie frontale et la partie périphérique. L'attention est portée principalement sur la zone face à l'auditeur, alors qu'il faut des transitoires importantes pour dévier l'attention vers la zone périphérique. Ainsi l'intelligibilité est augmentée lorsque les sources sont présentes dans le champ de vision.

Gil-Carjaval et al [GCSD16] ont montré, par la mise en situation de sujets dans le noir, puis en visuel, enfin dans une situation où la réverbération ne correspondait pas à la pièce, que l'externalisation est mieux notée dans le cas où les données sonores et visuelles sont en accord. Cependant, bien que l'externalisation soit mieux bien notée dans le cas où les sujets recevaient des informations de réverbération supplémentaires dissonantes avec celles à l'origine, cela ne change pas quand les sujets entendent un espace sonore différent de celui visuellement. Ils en ont déduit que la connaissance des réponses en fréquence d'une salle connues a un plus grand impact que les attentes relatives à la salle.

Larsson et al. [LVK02] ont déterminé que les sources sonores sont considérées plus larges avec l'utilisation d'un casque ou lorsque le sujet est directement dans la salle plutôt que face à des photos ou sans la présence d'indices visuels, ce qui peut confirmer la tendance des mixeurs à élargir le champ sonore lorsqu'ils mixent face à un petit écran, et à recentrer quand ils sont dans un auditorium.

Schutte et al. [SEW19] ont effectué un test pour déterminer l'influence de la vision sur la réverbération dans trois conditions : le son et l'image correspondent, ne correspondent pas, ou simplement le son est gardé. Ils n'ont pas trouvé d'influence significative, mais ils ne prenaient en compte qu'un aspect du son, alors que la spatialisation est une notion plus large (enveloppement, largeur, clarté, profondeur).

Salmon et al [SHEP20] ont cherché à déterminer l'influence de la vision sur différents espaces sonores. Lorsqu'il y a des conflits spatiaux entre le sonore et le visuel, le visuel l'emporte car la vue permet une meilleure localisation que l'écoute. L'hypothèse est que la vue de la pièce influence la perception sonore, en particulier si les deux ne concordent pas, basé sur les données des recherches précédentes. Afin de contrôler l'espace visuel perçu pour des espaces sonores donnés, ils ont eu recours à des casques VR qui présentaient des environnements visuels différents, et pour chacun de ces environnements était donnée une paire d'espaces sonores aux caractéristiques de réverbération et de taille différentes. Les résultats ne confirment pas l'hypothèse : il n'y a pas d'influence de la vision sur le jugement des paires.

1 – 8. Evaluation de la perception sonore

La perception sonore de l'espace est formée de plusieurs dimensions d'attributs [RG17]. Ils peuvent être arrangés dans une hiérarchie, avec un jugement intégratif de qualité en haut, et jugements d'attributs descriptifs individuels en bas. Selon le modèle de Letowski [L89], cet arbre peut être divisé largement en attributs de timbre et d'espace, les attributs d'espace référant aux trois dimensions du son que sont la localisation, largeur et distance, et les attributs de timbre référant aux aspects de coloration du son. Les effets de distorsion non-linéaires et le bruit sont parfois associés à ce groupe d'attributs de timbre.

Plus on monte cet arbre de hiérarchie, plus on parle de l'acceptabilité et de la pertinence du son pour un rôle donné et en relation avec un cadre de référence, alors que pour les niveaux les plus bas, chacun peut évaluer les attributs concernés séparés de toute référence de valeur. En d'autres termes, un haut niveau de jugement de qualité est une évaluation intégrante qui prend en compte tous les attributs inférieurs et pondère leur contribution. La nature de la référence, le contexte et la définition de la tâche gouvernent la façon selon laquelle l'auditeur décide de quels aspects du son devraient être pris en considération.

Bien que les chercheurs se sont concentrés sur l'analyse de la capacité des systèmes surround à créer des images fantômes parfaitement localisées et à reconstruire les fronts d'onde originaux correctement, d'autres facteurs subjectifs tels que la profondeur de l'image, sa largeur et l'enveloppement sont très fortement attachés à la préférence subjective dans les applications grand public. Ces facteurs sont plus durs à définir et mesurer, mais ils apparaissent néanmoins comme très importants pour définir la qualité globale (Mason, 1999).

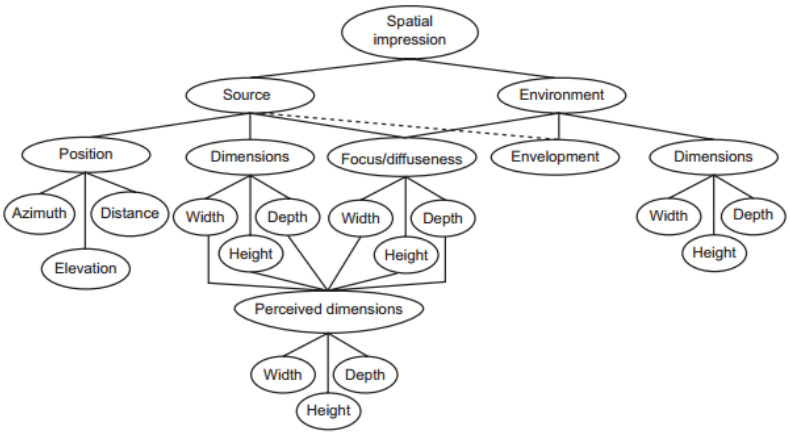


Figure 5 : Hiérarchie possible des attributs spatiaux dans une analyse subjective par Mason (1999). Source : Immersive Sound.

Si une vraie correspondance entre toutes les localisations de source était possible (ou seulement voulue/désirable), entre le milieu enregistrant et le milieu reproduisant, dans les trois dimensions et pour toutes les positions d'écoute possibles, alors il serait raisonnable de supposer que la capacité d'un système surround de créer de plausibles images fantômes de toutes les sources (en comprenant les réflexions) serait le seul prérequis pour la fidélité. Comme la vraie identité est rarement possible ou désirable, certains moyens de créer et contrôler les illusions adéquates pour les plus gros indices subjectifs pour le plaisir du consommateur peuvent être portés comme le but principal des techniques d'enregistrement et de diffusion. Cette attitude est particulièrement pertinente pour les applications grand public, mais ne sera pas forcément la bonne pour des simulateurs de vol par exemple.

Une observation intéressante sur les facteurs affectant la qualité globale de jugements du son surround était que la fidélité de timbre est considérablement plus importante que la fidélité spatiale [RZKB05]. Les auditeurs s'intéressent plus au timbre et à la couleur de la reproduction qu'à leurs caractéristiques spatiales. Les auditeurs naïfs ne remarquent que peu les aspects d'une localisation stéréophonique du son, et sont plus affectés par l'effet immersif des canaux surround ; ce sont seulement les auditeurs entrainés qui semblent apprécier les images fantômes correctes. La jonction entre les domaines spatiaux et fréquentiels (timbre) ne peut être ignorée, comme chacun de ces domaines influence la perception de l'autre [CBR14].

2 – La technologie binaurale

Le terme « binaural » renvoie dans un premier temps à la caractéristique humaine (et plus largement animale) de déterminer des sources sonores à l'aide de ses deux oreilles. Cependant, cette appellation relève maintenant d'une certaine technologie qui consiste à recréer un espace naturel d'écoute en trois dimensions au casque, à distinguer d'une simple stéréophonie. Cette technologie vise l'immersion, c'est-à-dire à reproduire le champ sonore naturel dans lequel l'auditeur évolue, à l'aide d'une compréhension poussée de ses capacités auditives.

2 – 1. Le principe

La stéréophonie du casque ne permet pas une écoute binaurale malgré l'excitation des deux tympans puisqu'il manque des données cruciales afin de repérer les positions des différentes sources : les fameuses HRTF. En effet, les haut-parleurs du casque sont collés au pavillon et n'ont donc pas assez de recul pour pourvoir les réflexions nécessaires sur le corps et l'oreille externe : le filtrage permettant la localisation est inexistant. De plus, les haut-parleurs sont situés sur le côté du crâne de manière permanente et ne permettent pas de localisation grâce aux ITD. La spatialisation (ou panoramique) se fait traditionnellement par une variation d'intensité d'une source entre les deux canaux, en amont de la diffusion, ce qui crée une forme artificielle d'ILD. Cela ne suffit malheureusement pas pour recréer une sensation d'immersion dans un espace 3D.

Comme le casque fait fi de toute physionomie externe pour diffuser le son directement dans le conduit auditif, il suffit simplement de reproduire tout le filtrage effectué par le corps lors de la réception d'un son entendu en champ libre. Un signal capté à l'entrée des deux conduits qui viendrait d'une source à une certaine distance rejoué au casque directement à l'endroit de la prise aurait alors pris en compte les HRTF et sonnerait au casque tel qu'il aurait pu sonner à l'auditeur originellement. Il est possible de recréer une image convaincante à l'aide de haut-parleurs mais le casque reste une solution privilégiée pour la maîtrise complète des signaux qui sont envoyés et qui sont reçus parfaitement : pas de diaphonie, pas d'acoustique de salle parasite, *sweet spot* constant, possibilité d'être totalement isolé des bruits ambiants parasites (dans le cas des casques fermés).

Cette technique d'appliquer le filtrage des HRTF sur un signal s'appelle une **convolution** (par les HRTF). Plusieurs cas de figures sont possibles pour obtenir des HRTF.

2 – 1. 1. Le binaural natif

Le binaural natif est la technique qui semble des plus évidentes : capter un signal contenant toutes les fréquences depuis une ou plusieurs positions depuis les canaux auditifs, ce qui donne un enregistrement binaural. Le résultat de ces enregistrements est

alors un fichier audio composé de deux canaux représentant l'oreille gauche et l'oreille droite. On peut utiliser pour ce faire une tête binaurale, c'est-à-dire une tête artificielle très proche de l'anatomie humaine (torse, épaules, tête, pavillons) avec un microphone omnidirectionnel par conduit auditif.

Ces têtes binaurales existent depuis les années 30, la première étant « Oscar » d'AT&T de 1932, et leur fonctionnement n'a pas beaucoup changé depuis, si ce n'est qu'au lieu d'être placés devant les oreilles, les microphones sont maintenant placés à l'intérieur. Les plus connues sont les têtes Neumann, qui n'ont malheureusement pas de torse, au nombre de 3 :

- La Neumann KU80 contenant des KM83 modifiés et égalisée pour une réponse linéaire en champ direct à 0°, sortie en 1973
- La Neumann KU81i contenant également des KM83 modifiés et égalisée pour une réponse linéaire en champ diffus, sortie en 1982
- La Neumann KU100, très utilisée aujourd'hui bien que sortie en 1992, contenant des micros KM100 et égalisée pour une réponse linéaire en champ diffus, avec filtres coupe-bas à 40Hz et 150 Hz ainsi qu'un pad de 10dB

Il existe également le mannequin pourvu d'un torse et d'épaules détaillés appelé KEMAR (*Knowles Electronic Manikin for Acoustic Research*), utilisé pour un grand nombre de prises de HRTF dans le domaine public [GM94] [XZY13] [WRS11].

Une autre technique consiste à enregistrer à l'aide d'un sujet humain, mais le principe reste le même : les microphones sont posés à l'entrée du pavillon. Bien que cette technique permette une individualisation parfaite des HRTF, elle apporte également un inconvénient qui est l'impossibilité d'immobiliser totalement le sujet et apporte alors des micromouvements (respiration, mouvements imperceptibles du crâne) qui peuvent fausser la perception spatiale.

En revanche, ce qui a été enregistré en tant que binaural natif est immuable, on ne peut pas séparer le signal s'il est complexe (un enregistrement musical par exemple) des HRTF, il est fait pour être écouté tel quel, à une position et pour une tête données.

Il existe cependant une méthode dite MTB pour *Motion-Tracked Binaural* qui enregistre selon plusieurs angles afin de garder l'auditeur au centre de la scène lorsqu'il bougera la tête lors de l'écoute. Plusieurs microphones pourvus de pavillons sont disposés le long d'une surface sphérique, ce qui permet de sélectionner le signal provenant du microphone le plus proche de la position de l'auditeur enregistré grâce à un head-tracker. Si l'auditeur se trouve entre deux positions, le signal des deux microphones sera alors interpolé. La localisation en binaural, au même titre que celle naturelle, est plus affinée et ainsi réaliste si un head-tracking est possible, cela étant dû aux micromouvements de la tête [H17].

2 – 1. 2. La synthèse binaurale

La synthèse binaurale est plus souple que le binaural natif puisqu'elle sépare l'aspect binaural (indices de localisation, HRTF) du contenu enregistré qui peut être une prise monophonique, multicanale ou même ambisonique. Il est alors possible de choisir des HRTF pour convoluer le signal, ou le **binauraliser**, comme on choisirait un filtre à appliquer.

Pour enregistrer les HRTF, la méthode est la même que pour enregistrer du binaural natif (avec une tête artificielle ou humaine), à la condition que le signal puisse être analysé et déconvolué par la suite. Comme le signal de test envoyé est connu, il est alors possible de le séparer du signal enregistré pour ne garder que le filtrage, par une analyse de Fourier. Nous avons vu que les HRTF ne sont pas les mêmes pour une direction donnée, il est alors nécessaire d'enregistrer ces signaux test venant de plusieurs directions et angulations possibles, discrétisés selon un angle choisi (cela peut aller de 2° à 30° en azimut et de 4° à 30° en élévation, mais typiquement ce sera 5° à 15°). Ils sont enregistrés dans des salles anéchoïques afin de ne garder que le filtrage d'un signal direct et de ne pas être parasité par la réponse en fréquence de la salle, mais il est également possible d'effectuer ces mesures dans une salle suffisamment grande pour que le signal indirect n'interfère pas avec le signal direct ou que les réflexions puissent être retirées par la suite [ADT01].

Une autre méthode consiste à décoder des données en ambisonique pour permettre une synthèse binaurale [NSMH03]. L'ambisonique est envoyé vers des haut-parleurs virtuels puis les signaux convolués avec les HRTF correspondant à leur position dans l'espace. Le head-tracking est déduit des données de matrices de rotation dans le domaine ambisonique. Cependant, c'est une technique qui demande beaucoup de ressources de calcul.

Comme nous l'avons vu, les HRTF sont spécifiques à chaque individu, on fait alors la distinction entre les HRTF dits individualisées et les HRTF génériques. Ces derniers peuvent être utilisés sans qu'il n'y ait de grands problèmes de localisation, comme l'a montré une étude de Wenzel et al [WAK93] en comparant des sources écoutées en champ libre et des sources virtuelles filtrées par des HRTF génériques, ce qui donne des résultats similaires malgré une confusion avant/arrière et haut/bas. Mais il y a également des études qui montrent des effets différents entre l'utilisation de HRTF individualisées ou génériques, comme un sentiment accru de présence [VLV04]. Dans une étude qui comparait des écoutes en champ libre, sans enregistrements, et des écoutes d'enregistrements avec des têtes humaines et des têtes artificielles, la justesse de localisation était moins bonne dans le cas des enregistrements, et au sein des enregistrements, moins bonne encore pour ceux faits avec la tête artificielle [MOC01].

Cependant, il est ardu sinon impossible d'obtenir les HRTF individualisées de tout le monde, comme c'est un processus qui demande du temps et des moyens lourds et coûteux (chambre anéchoïque, microphones, prise à chaque angle). Face à cette difficulté, il existe également une méthode pour sélectionner efficacement des HRTF dans une base de données [SK10].

Les BRIR (*Binaural Room Impulse Response*) sont des cas particuliers d'HRTF enregistrées dans une salle donnée. En plus du filtrage de l'appareil auditif est ajouté celui de la configuration de la salle, avec toutes ses réflexions et la nature de sa réverbération. La BRS, ou Binaural Room Scanning, est la technique qui consiste à enregistrer la réponse impulsionnelle d'une salle avec une tête binaurale artificielle [MFT99]. Cette technique peut être utilisée pour substituer une reproduction binaurale d'une salle à la salle-même. En effet, l'étude de Todd Welti et Sean Olive [WO12] met en œuvre des tests perceptifs où les sujets sont placés in situ (dans une voiture) pour écouter avec le système original et la reproduction binaurale au casque de ce même lieu. Le système binaural utilisé est le système BRS:AB, c'est-à-dire un enregistrement BRS mais où les programmes sont directement enregistrés au lieu d'une volonté de convoluer les programmes à la lecture. Le test consistait en deux écoutes de quatre programmes musicaux avec quatre effets appliqués qui touchent les caractéristiques spatiales de ces programmes : non modifié, biais vers la droite, augmentation de la corrélation, décalage de phase. La question était alors de savoir si la méthode BRS:AB donne lieu à des préférences identiques que celles données dans la condition in-situ. Les résultats confirment l'hypothèse que la reproduction binaurale peut être substituée à la situation in situ. Ils rejoignent une étude faite auparavant par les mêmes, pour cette fois un système BRS classique [OWM07]. On peut alors justifier l'utilisation de cette technique pour créer un espace cohérent et fiable en termes de reproduction spatiale. Il faut cependant prendre en compte que les deux écoutes se sont faites dans le même lieu.

2 – 2. L'immersion

Les dispositifs sonores immersifs sont de plus en plus nombreux et de plus en plus accessibles, mais la définition même de l'immersion reste un enjeu majeur de ces technologies. Qu'est-ce qui rend ces dernières si attractives par rapport à une simple stéréophonie ? De nombreuses études se sont penchées sur cette question cruciale, qui dépasse les cadres technique et psychologique.

2 – 2. 1. Définition

Le terme immersion est devenu très important sur ces dernières années dans le domaine du son, et est devenu synonyme de réalisme, naturel, présence, ou enveloppement [ABBF20]. Cependant ces termes désignent tous des concepts différents, et il convient alors de les séparer pour éviter toute confusion. Murray a d'abord décrit le terme comme suit : « L'immersion est un terme métaphorique dérivé de l'expérience physique d'être submergé par l'eau. Nous recherchons la même sensation d'une

expérience psychologiquement immersive comme nous le faisons en plongeant dans l'océan ou dans une piscine ; la sensation d'être complètement entouré par une autre réalité, aussi différente que l'eau l'est de l'air, qui capte toute notre attention, notre appareil perceptif entier. » Cette définition peut s'appliquer dans des cas qui ne recourent pas à la technologie à visée immersive, comme la lecture par exemple, où les ressorts narratifs et de style suffisent à donner cette impression, ce qui relève alors de l'état psychologique du lecteur. Mais dans le cadre des systèmes immersifs, il y a également un penchant technologique qui peut être mesuré objectivement.

Psychologiquement, l'immersion est soumise à trois facteurs : l'immersion perceptuelle, l'absorption dans la narration et l'absorption face au défi. L'immersion perceptuelle peut être mesurée par le rapport entre les sens excités par des *inputs* chez l'utilisateur et le degré de fermeture aux autres *inputs* de l'environnement. En d'autres termes, la capacité de l'utilisateur à se concentrer sur certains stimuli et ignorer les autres. Cela peut être atteint en bloquant toute interaction avec l'environnement externe au stimulus et en augmentant son pouvoir sensoriel. L'absorption dans la narration ou face au défi sont des notions qui se retrouvent notamment dans le jeu vidéo pour expliquer la capacité d'un joueur à se plonger dans un jeu même si les stimuli sont peu puissants, et relèvent ainsi d'autres codes pour garder son attention pleinement.

Une immersion purement définie par des critères objectifs sur des systèmes technologiques ne fait pas consensus, puisqu'elle dépendrait largement de son pendant psychologique. Ce n'est pas simplement en augmentant le nombre de canaux qu'un système devient fatalement plus immersif (voir le binaural), même si plus de données sensorielles sont produites. Mais cela va de pair avec la capacité de l'individu à se concentrer sur un stimulus au détriment des informations sensorielles environnementales. On peut alors parler de conditions ou propriétés technologiques qui faciliteront une expérience immersive.

2 – 2. 2. L'externalisation

L'externalisation est la capacité à situer une source sonore à l'extérieur de notre tête, ce qui est à opposer à l'internalisation, capacité à situer une source dans notre tête [BBLMK20]. Ces deux phénomènes correspondent à des situations différentes, l'une à l'écoute en champ libre et l'autre à une écoute stéréophonique typique au casque. S'il y a volonté de créer de l'immersion à travers le casque, il faut dépasser absolument tout sentiment d'internalisation.

Sans être forcément être indiscernable d'une écoute dite naturelle ou complètement réaliste, pour permettre l'externalisation, une écoute se doit de contenir les indices physiques qui permettent de créer une impression en accord avec l'expérience de l'auditeur. La non-satisfaction de cette impression amène alors à l'internalisation. Les sources doivent être reconnues dans un espace distal, c'est-à-dire hors de soi (on parle

aussi d'attribution distale). Ainsi les sources sont donc localisables à une certaine distance de soi, bien que la simple notion de distance ne soit pas suffisante puisqu'il est possible d'attribuer différentes distances au sein de sa tête, et donc de l'intégrer dans l'internalisation. Cependant la perception des sons externalisés montre une augmentation de l'activité dans la planum temporale du cortex cérébral [CCA13] [HGF03] impliquée dans la perception de la distance [KHB12]. Le cerveau gardant en mémoire des combinaisons de données binaurales des différents indices de localisation, une combinaison invraisemblable de ces indices amène à une internalisation du son. L'externalisation dépend alors beaucoup du binaural alors que la distance peut être perçue monauralement, cependant, des études montrent que perception et jugement de distance sont possibles malgré une faible externalisation [BL16] [KHB12], et une mauvaise externalisation amène des problèmes dans l'estimation de la distance [CSB13] [CBB18] [HW96] [HWD17].

3 – Les usages en cours

3 – 1. Techniques en cinéma

L'histoire du son au cinéma est jalonnée d'innovations techniques, il faut alors savoir s'il est possible que le binaural soit une étape plausible dans l'évolution du workflow audiovisuel.

3 – 1. 1. Le multicanal

Le terme multicanal recouvre plusieurs utilisations et techniques déployées dans diverses situations, mais ici nous parlerons des utilisations rencontrées le plus souvent en cinéma [H90]. Dans les années 70, Dolby s'impose dans le domaine du cinéma avec un système de réduction de bruit et d'équalisation de salle. Un ensemble de standards voit le jour, qui permet des transferts consistants entre n'importe quel auditorium et n'importe quelle salle de cinéma.

Les standards concernent différents paramètres acoustiques. La plage de fréquences et la réponse en fréquence à certaines positions d'écoute est définie par la norme ISO 2969 (dernière version date de 2015), qui vise un niveau perçu et une réponse en fréquence constante entre installations et emplacements dans les salles, pour une courbe obtenue appelée X (*X-curve*). Le niveau est normalisé à 85dB SPL. L'emplacement des haut-parleurs et le nombre de canaux est souvent le même, 5.1 ou 7.1, ce qui laisse une constante de 3 haut-parleurs à l'avant, un ou plusieurs subwoofers, et les haut-parleurs arrière arrangés pour avoir un champ uniforme. Le bruit de fond et le temps de réverbération sont contrôlés pour avoir la meilleure intelligibilité possible.

Linda Gedemer [G15][G16] a procédé à des tests perceptifs pour déterminer la qualité des courbes SMPTE et ISO (la fameuse courbe X). En effet, il n'y a jamais eu d'analyse subjective de ces courbes à partir des années 60, seulement des mesures par instruments « objectifs », alors que des tests perceptifs furent pratiqués dans les années 40 pour estimer les systèmes de salles de cinéma. Afin de tester différentes courbes pour plusieurs salles sur un grand panel de sujets, elle a procédé à un test par BRS, dont l'efficacité a été prouvée [OWM07]. Il en a été conclu qu'il y avait une préférence pour une courbe en particulier, qui suivait globalement la courbe X mais présentait un gros apport en-dessous de 100Hz. L'apport des tests subjectifs pourrait permettre une meilleure manière d'aborder notre écoute en salle.

3 – 1. 2. Le Dolby Atmos

Dolby sort en 2012 le Dolby Atmos, qui est un système de mixage orienté objet. Le nombre de canaux peut être drastiquement augmenté, mais l'avantage de ce système est sa transposition à n'importe quel système. Dolby se targue de proposer une expérience

immersive supérieure à ce que peuvent proposer les systèmes surround existants. En effet, la particularité de l'Atmos par rapport aux autres configurations physiques de haut-parleurs est leur présence au plafond, qui veut émuler la dimension verticale oubliée auparavant. En tant que système de mixage orienté objet, les sons peuvent être effectivement séparés sur chaque haut-parleur plutôt que présents sur la rangée entière dans le cas du surround, selon des données de panoramique traitées en marge de l'audio.

Oramus et Neubauer [ON20] ont testé la perception spatiale entre des mixages classiques surround et des mixages orientés objet en Dolby Atmos. Les sujets sont placés dans une salle de cinéma équipée Atmos et doivent estimer la position de sons courts mixés selon différentes méthodes (5.1, 7.1 et Atmos) sur un graphique à deux axes en évaluant également leur confiance dans leur réponse. Bien qu'il ait été montré précédemment qu'un nombre accru de canaux améliore la sensation d'enveloppement et la finesse de localisation, le résultat de leur étude indique qu'il n'y a pas de différence en précision de localisation en fonction des formats, mais que l'assurance dans la réponse est plus élevée dans le cas du Dolby Atmos.

Le binaural pourrait être une solution adaptée pour écouter des mixages faits en Atmos car il permet de restituer la composante d'élévation, ce qui est difficilement reproductible dans une salle de montage équipée en 5.1. De plus, le coût d'un dispositif au casque est nettement avantageux par rapport à celui pour adapter une salle au système Dolby Atmos.

3 – 2. Un marché en expansion

3 – 2. 1. Un studio sans studio : utopie ou réalité ?

Dans le milieu professionnel, bien que le binaural pouvait être un choix de prise de son ou de diffusion, il est assez récent de considérer sérieusement le binaural comme une possibilité de monitoring. En dehors de Flux et de l'Ircam qui proposait des plug-ins de binauralisation, le marché s'est étendu aux alentours de 2015 avec une volonté de toucher le plus de techniciens possibles lorsque des entreprises qui orientent leurs produits vers le grand public ont sorti leur propres plug-ins ou hardware.

Waves NX

Figure 1 : Capture d'écran de Waves NX sur Reaper

En 2016, le constructeur de plug-ins Waves sort Waves Nx (ou Nx Virtual Mix Room) qui permet de simuler une pièce traitée acoustiquement avec une technologie de tracking par tracker bluetooth ou vidéo. Le plug-in peut gérer du surround (7.1, 7.0, 5.1, 5.0), de la mono, de la stéréo mais aussi de l'ambisonique d'ordre 1. Il fonctionne sur une synthèse binaurale qui permet une individualisation relative des HRTF par calcul des ITD et ILD selon les renseignements donnés par l'utilisateur (circonférence du crâne et arc inter-aural). Il n'est donc pas possible d'importer des HRTF, l'algorithme du plug-in s'occupant de les générer.

Le *room amount* permet de doser la quantité de réverbération de la salle, ou du moins de simuler une plus grande distance. Le center trim permet de gérer la quantité de centre envoyé dans l'algorithme, si bien qu'il est possible de l'isoler totalement pour des dialogues par exemple.

Steven Slate Audio VSX

Steven Slate Audio sort en octobre 2020 le casque VSX accompagné d'une application logicielle qui ensemble permettent de simuler différents lieux d'écoute et corrections acoustiques. Le fabricant utilise des algorithmes appelés BPM pour *Binaural Perception Modeling*, basés sur une technologie appelée Quadvolution par les développeurs de Scaeva qui a contribué à sa création.

Spatial Sound Card Pro

Cette application de New Audio Technology peut fonctionner en plug-in comme en standalone, ses utilisations étant multiples. En effet, cette carte son virtuelle peut être utilisée comme monitoring mais également en upmixer de diffusion pour différentes situations. Sa version pour les DAW (Digital Audio Stations) est Spatial Audio Designer

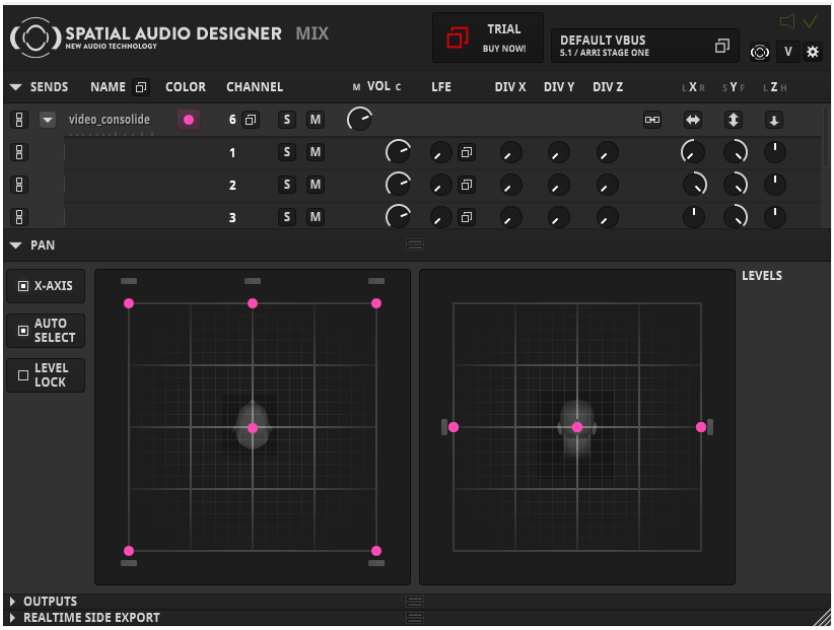


Figure 2 : Capture d'écran de Spatial Audio Designer sur Reaper

Le plug-in est composé de plusieurs parties, dont la partie Send qui doit se mettre en insert de la piste devant être traitée. Le plug-in présent sur le bus d'écoute reçoit alors les différents flux, et chaque canal peut être envoyé n'importe où dans le panner.

Sienna Acustica

Figure 3 : Capture d'écran de Sienna Acustica

Suite de plug-ins de simulation de salle avec un grand nombre de situations disponibles, Sienna Acustica sort en 2021 dans la lignée de Waves NX ou SoundID Reference. 3 plug-ins font partie de cette suite qui se décline ainsi :

- Sienna Reference qui permet de corriger les défauts d'égalisation du casque, en choisissant le modèle et en réglant la quantité d'effet appliquée (Dry/wet).

- Sienna Rooms, qui est la simulation de salle à proprement parler, permet donc de sélectionner une pièce conforme aux standards professionnels ou de choisir un autre support de diffusion comme seconde écoute. Des expansions permettant un plus grand choix de salles et de support sont prévues.

- Sienna Guru permet de gérer la configuration de la pièce plus en détail, en jouant sur des paramètres de taille, de volume, etc.

3 – 2. 2. Le cas particulier des simulations d'auditoriums cinéma

Les plug-ins présentés jusqu'ici ne sont pas forcément destinés uniquement aux professionnels du cinéma mais à un public plus large, voire seulement aux ingénieurs du son spécialisés dans la musique. Cependant le marché des processeurs binauraux existe également, et ce depuis longtemps malgré une implémentation tardive dans les auditoriums.

Tomlinson Holman's C41 MicroTheater

Dès 1996, Tom Holman, qui a travaillé 15 ans comme directeur technique chez Lucasfilm, puis a fondé THX dans les années 80, propose un système qui entre complètement dans l'évolution des pratiques pour les 20 ans qui suivront, à savoir proposer un système qui permet de pré-mixer dans la salle de montage.

Le système ne fonctionne pas en binaural mais s'adapte à l'acoustique des pièces par des mesures et des corrections apportées par le processeur C41 (« *Cinema-for-one* ») et propose une diffusion en 4.1, avec centre fantôme. L'effet stéréo est supprimé par une annulation de la diaphonie par un signal inverse de celui reçu par l'oreille opposée au haut-parleur par l'autre haut-parleur.

Ce système émerge d'un constat et de situations professionnelles où, avec l'arrivée des premiers stations audio-numériques en post-production, se posait la question de la place du son et en particulier du montage son, qui déplace des questions réservées jusque-là au mixage vers le montage. Une première pierre à l'édifice qui signera l'avènement du montage son tel qu'il est pratiqué de nos jours.

Lake Technology TheaterPhone ou Dolby Headphone

Le TheaterPhone des australiens de Lake Technology ou Dolby Headphone est une technologie du début des années 2000 présentée sous une forme hardware ou en plug-in pour Pro Tools version 5.1. Son application se veut destinée à la post-production audiovisuelle.

La version hardware (HSM6240) se présente comme un rack 1U avec entrées 5.1 en jack TRS et sorties RCA et TRS pour les casques à l'avant, avec possibilité de passer de -10dBV à +4dBu pour les entrées. Les contrôles sont à l'avant, avec trois simulations de salles au choix : DH1 étant une salle très mate, DH2 une large salle d'écoute, et DH3 une grande salle de cinéma ou auditorium. La version plug-in utilise un DSP entier dédié à Pro Tools.

Beyerdynamic's Headzone Pro

En 2007, Beyerdynamics propose un processeur hardware qui inclue un head-tracker pour la première fois, grâce à une technologie de trackers à ultrasons qui permettent de calculer en permanence l'angle de la tête de l'auditeur. Le choix des sources peut être analogique ou numérique (entrées asymétriques 5.1 ou Firewire) et le casque utilisé est une version modifiée du DT880 Pro HT.

Smyth Realiser A8 [SSCK10]

Le Smyth Realiser A8 est un processeur entièrement hardware qui permet l'enregistrement et la lecture de HRTF, puis de les appliquer à des signaux surround allant jusqu'au 7.1, en entrée numérique (HDMI) ou analogique (TRS). Cette machine à plusieurs milliers d'euros est très répandue dans les studios de post-production et est fournie avec ses microphones pour la prise ainsi que son head-tracker. Tout est prévu pour une utilisation sans avoir recours à des outils externes, le processeur pouvant également enregistrer la réponse en fréquence des casques utilisés couplés avec des BRIR. Il y a alors l'avantage de la personnalisation des BRIR et de la courbe d'égalisation du casque utilisé ainsi que la gestion des micromouvements (entre -30° et 30° azimut) grâce au head-tracker.

Figure 4 : Avant et arrière du Smyth Realiser A8

Chaque paire binaurale est convoluée avec un signal venant d'un canal, puis tous les signaux convolués sont sommés dans la paire binaurale. La convolution se fait à la lecture grâce aux différents angles interpolés des PRIR (*personalized room impulse*

responses). En ce qui concerne le niveau de sortie, la calibration se fait grâce aux signaux de test du Smyth ou du bruit rose envoyé par les enceintes et le casque : quand les niveaux sont considérés égaux, le niveau en SPL est inscrit dans le fichier. Il est également possible de gérer retards et délais qui peuvent être pris en compte au moment d'enregistrer dans l'auditorium, qui peuvent déjà les implémenter dans leur chaîne de diffusion.

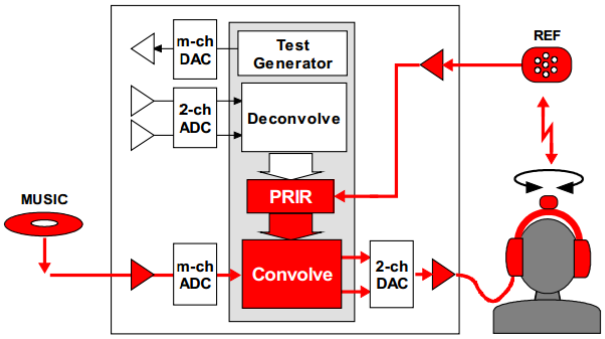


Figure 5 : Fonctionnement du SVS. Source : Smyth

3 – 3. Besoins et contraintes

Suite à la présentation des différentes solutions à la question de la reproduction binaurale de salle présentes sur le marché, nous pouvons en conclure que l'offre est présente et tend à se multiplier. Elle prend différentes formes, entièrement software (Waves NX, Sienna, Spatial Audio Designer), hybride (Steven Slate VSX) ou hardware (Smyth Realiser A8). Certaines formes sont pourtant préférables à d'autres, notamment l'ergonomie et l'aspect pratique d'un plug-in comparé à une machine telle que le Realiser A8, mais c'est pourtant ce dernier qui est retenu dans les studios, au regard de ses capacités.

Cyril Holtz est un mixeur français reconnu qui a pris l'habitude de travailler au casque sur du mixage à plusieurs en auditorium ou sur du prémixage. Son studio HAL a fait l'acquisition d'un Smyth Realiser A8 qui lui permet d'enregistrer et stocker les BRIR et réponses en fréquence des casques associés. Cependant, suite à une utilisation répétée de ce dispositif, plusieurs problèmes ont été soulevés :

- Les reproductions binaurales seraient très convaincantes dans le lieu où les BRIR auraient été prises, mais beaucoup moins si elles sont utilisées dans un autre contexte, ce qui est embêtant puisque l'intérêt réside dans la possibilité de se passer matériellement de l'auditorium. En effet, il y aurait une trop grande disparité entre ce qui est vu et ce qui est entendu, malgré l'adaptation rapide à une écoute binaurale, les repères visuels jouant un rôle immense dans l'appréhension de l'écoute. Des IR de salles plus petites sont plus tolérables.

Ce problème a un impact sur le mixage, qui est de réduire la quantité de réverbération appliquée sur les sources.

- L'absence d'un caisson de basses ne permet pas de travailler correctement en fréquence.
- La dynamique est souvent trop importante, dû à l'écran, trop proche et trop petit. Cela impacte également la panoramique.

Les autres problèmes soulevés par d'autres professionnels du son à l'image, Boris Chappelle, étaient de l'ordre de l'ergonomie relative au casque. Bien qu'ayant remarqué une diminution de la taille des auditoriums, il n'envisage pas un travail au casque à cause de l'inconfort que cela génère. Les monteurs parole américains ont pourtant recours au casque quand il s'agit des dialogues, et il est courant pour des mixeurs musique de vérifier leurs mixages au casque, dans une optique de double écoute. Mais pour des scènes chargées avec des éléments qui évoluent en permanence en matière, en espace et en temps, et pour des productions qui peuvent prendre plusieurs semaines de travail, il est compréhensible que le casque ne semble pas être une solution tout au long de la journée de travail.

Pierre Bézard [B13] s'était intéressé à l'utilisation du binaural dans le cadre de la post-production et de la diffusion en se basant sur un test perceptif fait uniquement au son (sans image) qui déterminait la qualité de transfert d'un mix 5.0 vers une réduction binaurale. Il en a conclu que le binaural est une solution satisfaisante en termes d'immersion et de plaisir d'écoute même chez des naïfs en la matière, et que l'individualisation des HRTF est primordiale. Nous faisons l'hypothèse que cette conclusion positive est un argument pour continuer dans cette voie, et nous amènerons la dimension visuelle dans les tests, qui a une influence évidente sur l'écoute, comme nous l'avons montré en partie 1 – 7.2.

Nous avons vu que la standardisation est une pratique qui touche toutes les caractéristiques du son à l'image en vue de diffusion (pas qu'en cinéma d'ailleurs, la norme R-128 en télévision est également largement respectée). Pour parer à cette éventualité, il faudrait prévoir une norme qui encadrerait les mixages en binaural, particulièrement le niveau d'écoute au casque, ainsi qu'un test pour choisir efficacement les BRIR ou HRTF.

En prenant toutes ces données en compte, de l'analyse des solutions présentes sur le marché à la présentation de la technologie binaurale, nous pouvons en déduire un dispositif qui soit quasiment optimal. Bien que le Smyth Realiser A8 soit très répandu dans les studios et la meilleure solution actuellement, car il prend en compte les mouvements de la tête, l'individualisation des HRTF ainsi que la réponse en fréquence des casques utilisés, il reste difficile d'utilisation. La télécommande et l'aspect général de la machine sont des technologies qui semblent archaïques au vu de tous les plug-ins puissants et ergonomiques. Une solution possible serait une technologie hybride entre la

contrainte hardware (prise de HRTF, head-tracking) et une partie software qui pourrait gérer la convolution. De plus, un des problèmes du Smyth est sa rigidité concernant sa banque de HRTF, puisque sont stockées uniquement les BRIR enregistrées grâce à la machine. Il pourrait être intéressant de pouvoir importer des BRIR d'autres banques.

Pour pallier le problème de la dissonance entre l'espace vu et entendu, un casque VR pourrait être une solution. Même si l'étude de Salmon et al. présentait des résultats indiquant que les différences de tailles entre espaces visuels en VR n'influençaient pas la perception spatiale sonore, la différence de mixage et de perception est empiriquement remarquée par les mixeurs et monteurs son en fonction de leur lieu de travail et de la taille de l'écran dans ce lieu. Mais le casque VR est probablement bien plus inconfortable qu'un casque tout court, qui l'est déjà suffisamment pour susciter la réticence des monteurs quant à son utilisation régulière. Dans le cadre d'une simple vérification de mix, ce pourrait être une solution intéressante, mais pour le travail en salle de montage, il n'y a malheureusement à ce jour aucun dispositif qui soit miraculeux.

Partie 2 : Partie expérimentale

1 - Principe

Nous avons vu que la demande et la technologie pour un outil permettant d'écouter un mixage in situ chez soi ou en salle de montage existent. Ce mémoire vise à la création d'un protocole qui permettrait de valider ce passage d'un mixage effectué en conditions « réelles », soit en auditorium, et un mixage effectué en binaural dans cet espace reproduit.

Nous choisirons de faire des prises de BRIR dans un auditorium de l'école pour la facilité du dispositif, sa fiabilité et sa flexibilité (possibilité d'individualiser les BRIR et ainsi proposer aux mixeurs un grand choix de BRIR). Malheureusement les contraintes de temps et de technologie ne nous permettent pas d'inclure un head-tracking qui aurait pu compenser les erreurs de localisation ainsi que celles de prises dues aux micromouvements de la tête.

Il a été décidé de recourir à des tests perceptifs comme méthode d'évaluation d'un transfert réussi ou non pour un même mixage réalisé dans deux situations différentes. Ce même mixage devra être fait sur des extraits tirés de films, avec l'image, l'influence de l'image sur le son étant importante et décisive dans un cas concret de mixage en post-production. Le montage serait également fait en amont et proposé à mixer selon des directives que je donnerais aux personnes qui mixeront telle une réalisatrice. Ainsi, la seule caractéristique variable pour un mixage donné et qui sera évaluée in fine sera la condition matérielle dans laquelle il aura été fait, c'est-à-dire en auditorium ou en salle de montage au casque en binaural. Le transfert sera évalué dans l'auditorium, c'est-à-dire que les différents mixages seront jugés dans le lieu où ils sont censés avoir été mixés, l'un in situ et l'autre en reproduction binaurale.

Si le transfert est concluant, peu de différences devraient être notées entre les deux mixages, ou bien un consensus devrait être atteint sur une viabilité égale des deux en vue d'une diffusion commerciale ou dans un cadre de travail. Si toutefois des différences nettes et systématiques sont remarquées lors de l'analyse statistique, elles permettront de définir les écueils ou caractéristiques particulières d'un mixage en binaural par rapport à un mixage traditionnel.

Afin d'établir des tendances il faut avoir un grand nombre de mixages, et pour retirer la variable d'un mixeur en proposer plusieurs également. Ainsi, j'ai pu avoir un mixeur (D.K.) et une mixeuse (A.L.) qui ont chacun fait les sept mêmes extraits, sélectionnés sur 4 films différents, dans deux situations différentes (in situ et casque), et qui ont eu la possibilité d'utiliser leurs propres HRTF.

2 – Préparation des mixages

2 - 1. Enregistrement des BRIRs

Les BRIRs ont été enregistrées dans l'auditorium en une journée, car les prises en elles-mêmes sont très courtes (une minute par prise). Les mixeurs ont eu leurs BRIRs enregistrées, ainsi que les miennes.

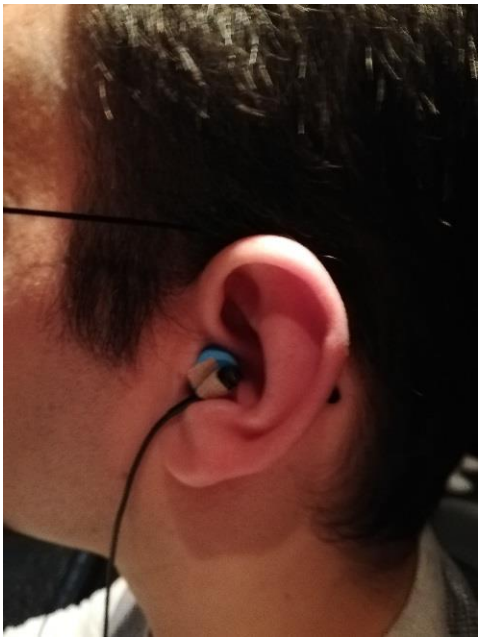


Figure 1 : Placement du 4060 dans l'oreille du mixeur 1 (D.K.)

Figure 2 : Emplacement de la prise, au sweet spot, dans l'axe de l'écran

Le dispositif était donc composé d'une paire de DPA 4060 (avec bonnettes) scotchés aux oreilles des mixeurs, les aigus orientés vers l'extérieur. La carte son utilisée pour l'enregistrement était une Focusrite Scarlett 2i2 reliée à un Macbook. Les fichiers ont été enregistrés sur Reaper en 48kHz et 24 bits.

Les sujets ont été placés au *sweet spot* ou à l'emplacement habituel pour le mixage, au niveau du deuxième bac de faders et dans l'axe de l'écran, le regard visant le centre de ce dernier. Nous avons fait attention à ce que les sujets gardent la tête la plus droite possible, dans une posture en alignement avec le dos.

Le fichier de test servant pour la déconvolution est un *sweep* de 5 secondes par canal, donc 5 sweeps en tout, chacun séparé de 5 secondes de silence. Afin d'avoir une plus grande marge de manœuvre, 5 prises ont été faites par tête. On se retrouvait donc à la fin avec 25 prises, mais ce sont les trois dernières de chaque qui ont été gardées pour créer les BRIRs à proprement parler.

La déconvolution a été faite par François Salmon, ingénieur spécialisé dans le développement de plug-ins, qui a utilisé une transformée inverse de Fourier sur les sweeps (a utilisé un sweep inversé). L'intégration en SOFA dans le plug-in Binauralizer Studio a été faite par Charles Verron, fondateur de Noise Makers.

2 – 2. Choix des extraits

Les extraits ont été choisis pour leurs variétés de sons et de situations : cinq des sept comportent des dialogues allant de un à trois personnages, les deux autres mettent en avant les effets et bruitages, deux scènes utilisent de la musique (extra-diégétique) et trois sont composés de plus d'un plan (champ/contre-champ principalement). La session qui regroupe les extraits utilisés pour les tests (1, 2, 4, 5, 7) avec tous les mixages est disponible au lien suivant : https://drive.google.com/drive/folders/1X4oMxpV6oEqkmHOujwvBuhwBlTtPaSe0?usp=sharing.

1 : Extrait du film "L'Amazonie brûle encore" de Helio Pu (Fémis, 2021)

Cet extrait comporte deux espaces dans le plan, un dans une télévision et l'autre la salle dans laquelle elle se trouve. Un personnage entre par la droite et explose l'écran avec un marteau. La vidéo diffusée par la télévision doit être traitée comme telle car elle a été incrustée par la suite ; il y a ainsi des directs (télé), des ambiances (télé et salle), des bruitages (télé et salle) ainsi que des FX pour l'explosion de l'écran. Choisi pour la différence brutale de dynamique, la gestion des deux espaces simultanés, le placement des sources dans l'espace intérieur.

2 : Extrait du film "L'Amazonie brûle encore" de Helio Pu (Fémis, 2021)

Cet extrait comporte très peu de sources (direct, ambiances) et est centré autour d'une action dans un lieu à l'acoustique très particulière qui doit être amplifiée par le mixage et l'apport d'une réverbération. Il y a également une entrée dans le champ de droite au centre du personnage, à spatialiser donc.

3 : Extrait du film "L'Amazonie brûle encore" de Helio Pu (Fémis, 2021)

Cet extrait comporte du direct, de la musique, des ambiances et une grande quantité de FX. L'action se déroule dans un endroit irréel à définir acoustiquement par une réverbération très large, et les divers FX permettent une spatialisation marquée très artificielle (nappes allant d'arrière en avant ou d'avant en arrière, largeur ou étroitesse des sons, etc.).

4 : Extrait du film "Cool kids" de Arthur Chrisp (Louis-Lumière, 2019)

Cet extrait est fait d'un champ et d'un contrechamp avec trois personnages, deux qui ne parlent que sur leur champ et un parlant en continu même fors champ. Il convient donc de travailler l'acoustique de cet intérieur en fonction du champ et du hors champ et faire ressortir les dialogues malgré le monologue du personnage. Les ambiances sont nombreuses et à spatialiser.

5 : Extrait du film "Cool kids" de Arthur Chrisp (Louis-Lumière, 2019)

Cet extrait est fait d'un plan qui suit un personnage s'avançant vers une source de dialogues. Il faut donc jouer ce rapprochement de la source par un dosage entre perche, HF et réverbération. De plus, la musique extra-diégétique doit être mixée en fonction des dialogues des personnages.

6 : Extrait du film "Les rituels" de Manon Sabatier (Louis-Lumière, 2019)

Cet extrait est fait de plusieurs plans mais reste centré sur un personnage qui parle aux autres dans un lieu réverbérant (carrelage) et assez étroit. Le placement de la voix dans le lieu est alors primordial, ainsi qu'accorder les ambiances et effets à ce même lieu.

7 : Extrait du film "La fatigue" de Louise Giboulot (Louis-Lumière, 2019)

Cet extrait comporte plusieurs lieux avec des ambiances très différentes et un traitement des voix qui l'est tout autant. Deux personnages indiquent par leurs mouvements de tête la provenance d'une source sonore émergente, filtrée par une chaîne hi-fi, à mettre en lien avec le monologue d'un troisième personnage dans un lieu différent du leur.

2 – 3. Déroulement des mixages

Les mixages se sont déroulés en auditorium (in situ) dans un premier temps puis en salle de montage (au casque, en binaural). Les monteurs ont mixé durant trois heures environ les sept extraits provenant de quatre films différents afin de faire varier autant que possible les sources.

Les mixages seront désignés par les noms A1, A2, B1 et B2, respectivement le mixage en auditorium du monteur A, le mixage au casque du monteur A, le mixage en auditorium de la monteuse B, le mixage au casque de la monteuse B. L'auditorium est celui de l'école Louis-Lumière et est équipé en enceintes Meyer, d'un Pro Tools 12.7 ainsi que d'une S5 de Euphonix.

Le plug-in utilisé pour permettre un monitoring binaural en binauralisant l'écoute est le plug-in Binauralizer Studio de Noise Makers. Ce plug-in est facile d'utilisation et s'insère sur le master d'écoute. Il est possible de choisir une configuration binaurale parmi les nombreuses proposées, et d'importer ses propres HRTF dans des fichiers SOFA[1], ainsi que de choisir le format de diffusion et la présence d'un passe-bas pour le LFE.

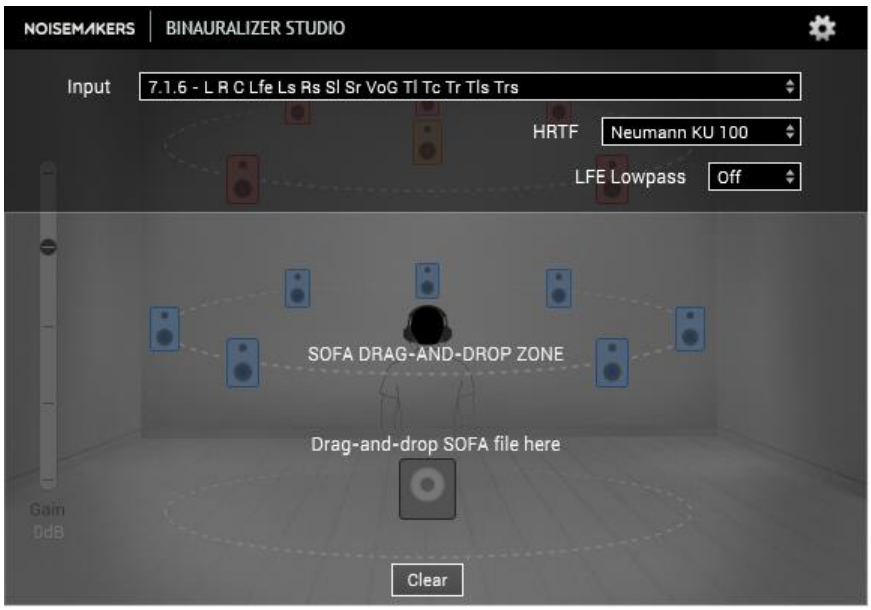


Figure 3 : Capture d'écran de Binauralizer Studio sur Reaper

2 – 3. 1. Déroulement du mix A1 en auditorium

Ce mix a demandé deux heures de préparation car étant le premier des deux à être effectué dans cette salle et a été décisif dans l'orientation du protocole. L'organisation de la session a été faite par le mixeur suivant son habitude de travail à partir de la base que

[1] SOFA : Spatially Oriented Format for Acoustics, ou format de fichier qui contient des HRTF ou des BRIR, standardisé par l'AES (AES69-2015).

je proposais, assez différente d'une session de travail habituelle car comportant moins de sources, donc de pistes, et une absence de quinconce car les extraits étaient très courts.

Les extraits ont été mixés selon l'ordre suivant : 4, 5, 6, 7, 1, 2, 3. Les extraits qui demandaient le plus de travail ont donc été faits en premier, ce qui atténue la fatigue sur les derniers extraits. Des pauses ont été faites tous les heures et demie environ.

Les plug-ins utilisés sont les Channel strip pour tout ce qui relève de la compression, des filtres et des EQ, ainsi que des Altiverb (mono pour les directs et les bruitages, stéréo pour tous les stems et quad pour la musique). Un plug-in Lo-fi (Avid) a été utilisé temporairement pour un effet radio dans l'extrait 7.

Les niveaux des auxiliaires de réverbération sont au niveau nominal (à 0) sur l'ensemble de la session, seuls varient les niveaux d'envoi des pistes vers les auxiliaires ainsi que le « mute » des envois.

2 – 3. 2. Déroulement du mix A2 au casque

Ce mixage s'est déroulé sur mon poste de travail chez moi car cela nécessitait une connexion internet et une version récente de Pro Tools pour le plug-in Binauralizer Studio. Il n'y avait pas de contrôleur midi ou Eucon, seulement un clavier et une souris pour toutes les automations. La carte son était une M-audio M-track 2x2 et les écrans 24 pouces au nombre de deux. Le Pro Tools utilisé était Ultimate 2021.

Le choix du casque s'est fait par l'habitude d'utilisation du HD-25 par le mixeur, et parce que le casque est fermé. Pour le choix des BRIR, le mixeur a pu tester les quatre disponibles (les siennes, celles de l'autre mixeuse, les miennes, celles de la Neumann KU-100) sur un son test (extrait d'interview radio, sans image). Il a choisi les siennes sur des critères d'espace et de précision d'emplacement des sources.

De manière générale, tout a été bien plus rapide, la session avait déjà été configurée au préalable et le mixeur connaissait les scènes. Les mixages enregistrés au préalable de la première session ont servi de référence en termes de niveau d'écoute et de rappel sur la direction artistique des mixages, au risque de biaiser le travail sur la présente session. Une comparaison était faite systématiquement en fin de mixage à chaque extrait, une fois le mixage validé, et aucune modification n'était apportée par la suite.

Son retour sur l'expérience est de nature aussi bien technique qu'ergonomique : les directs lui semblaient chargés ou fouillis (notamment sur l'extrait 4), ce qui peut amener à diminuer l'apport en réverbération. Une comparaison entre le mixage écouté par le bus binaural et par le bus de downmix "simple" permet de se rendre compte de l'acoustique apportée par la binauralisation et ainsi noter une différence importante entre les deux, la

préférence étant pour l'écoute sans binauralisation. Les placements dans l'espace, en particulier les panoramiques, semblent flous, il n'y a pas de réelle certitude spatiale sur ce qui est en train d'être mixé quand on se projette dans la salle de cinéma (car il faut compenser en permanence en recentrant les éléments, mais cela peut dépendre aussi de la taille de l'écran utilisé). Il en a donc déduit que c'était un bon outil pour vérifier ou réécouter un montage.

2 – 3. 3. Déroulement du mix B1 en auditorium

La session avait déjà été préparée au préalable par le mixeur précédent, et pour des raisons de temps est devenue la base de travail malgré des habitudes entre les deux qui n'étaient pas les mêmes.

La mixeuse B a moins l'habitude du mixage en auditorium et est plus habituée du prémixage. Cela a pu influencer le temps passé et la fatigue ressentie lors du mixage.

Les EQ utilisés étaient les Pro-Q 2, les plug-ins de réverbération des Altiverb 7 stéréo ou 4,0. Les auxiliaires de réverbération étaient à 0, montés sur l'ensemble de la séquence vers -10, avec des envois variables.

2 – 3. 4. Déroulement du mix B2 au casque

La session s'est déroulée dans une salle de montage chez HAL, équipé d'un Pro Tools Ultimate 2020.9, d'un HD Omni et d'un écran de télévision pour diffuser l'image. Cependant, il n'y avait pas de surface de contrôle, ce qui aurait pu faciliter grandement certaines passes et a pu gêner.

Le casque utilisé était un DT-770 apporté par la mixeuse qui le connaît donc très bien et reconnaît son confort physique. Cependant, à l'écoute il présente un "lissage" fréquentiel (et donc spatial après binauralisation) qui rend plus flou les emplacements des sources et crée un englobement général. Ces défauts sont devenus bien plus flagrants une fois la comparaison faite avec le HD-25, qui rend plus nette la position des sources dans l'espace. Les BRIRs choisis étaient les miens (Grivelet) car plus large à l'avant, donnant un espace plus cohérent.

Les extraits ont été mixés dans l'ordre 4, 5, 6, 7, 1, 2, 3, au vu des difficultés qui ont été rencontrées lors de la première session.

Les premiers retours sur le binaural (avant de mixer) portaient sur l'espace et le timbre : l'espace semble étrange (décorrélé de l'image et du lieu), les voix sonnent pincées et agressives. Les aigus et les graves manquent de précision.

Après le premier extrait, la mixeuse a remarqué qu'elle se concentrait plus sur l'écoute que sur l'image (ne la regardait qu'après avoir mixé en général) pour cause de décorrélation entre image et espace sonore. Les voix résonnent beaucoup dans ce qui donne l'impression d'être un espace restreint et réduit le besoin apparent de réverbération. Le centre a l'air flottant, les voix semblent assez larges par rapport au centre. Après comparaison avec le mix enregistré en auditorium, quelques différences légères sont relevées, notamment sur les suivis de niveaux (ce qui peut s'expliquer par l'absence de surface de contrôle dans cette deuxième session).

Lors du deuxième extrait, l'utilisation d'une réverbération 5.0 a semblé plus cohérente du fait de l'englobement "flou" dû à la binauralisation.

A la fin de la session, la mixeuse en a conclu que les versions des mix faites en auditorium sont plus détaillées que celles faites en binaural et que le visuel (lumière restées allumées, écran de télévision) joue aussi beaucoup sur l'impression de réalisme et d'immersion. De plus, le fait de porter un casque (même plutôt confortable) entraîne une fatigue plus rapide. Le temps d'adaptation au dispositif est cependant assez court (au bout d'un extrait). Les différences sont plus ou moins grandes en fonction de la matière sonore et du type de scène ; s'il y a des détails en placement acoustiques, en niveau ou en fréquence, le casque n'est pas forcément conseillé. Il y avait également des problèmes de phase lorsque les panoramiques étaient trop éloignées des bords.

3 – Déroulement des tests perceptifs

3 – 1. Protocole

A la fin des mixages effectués avec le mixeur et la mixeuse, nous nous retrouvons avec sept extraits, chacun comprenant quatre mixages (A1, A2, B1, B2). Le but est de les faire écouter à un panel de sujets experts avec l'objectif de déterminer si le transfert entre l'auditorium et sa version binaurale est acceptable ou non. De simples critères objectifs déterminés par des mesures avec analyseur de spectre ou de sessions ne sont pas suffisants, surtout dans un cadre où l'image et la sensibilité jouent un rôle primordial.

Le test perceptif consiste en une diffusion aléatoire des extraits et de l'ordre des mixages pour chaque sujet, divisé en deux sessions d'écoute comportant les mêmes extraits, dans un ordre différent en fonction de la session. Le sujet est placé au *sweet spot*, soit à 6m de l'enceinte centrale, derrière la console. Je gère le lancement des extraits et le passage d'un mixage à l'autre pour chaque extrait. Les réponses se font sur une feuille avec un stylo.

La première session porte sur une appréciation globale par extrait, en demandant au sujet « Comment évalueriez-vous votre préférence entre A et B ? » (tour à tour casque ou auditorium), entourant sa réponse sur une échelle à 11 points, graduée sans chiffres. Les extrêmes se veulent des différences nettes, ainsi la réponse est formulée ainsi au-dessus des graduations 0 et 11 : respectivement « A nettement préféré à B » et « B nettement préféré à A ». *(voir annexe 1)*

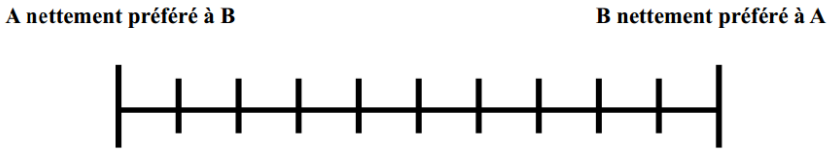


Figure 1 : Echelle de réponse utilisée dans la première session

La deuxième session se veut plus analytique, avec la même échelle de gradation mais des questions différentes *(voir annexe 2)* :

- Comment évalueriez-vous la qualité du timbre entre A et B ? (A nettement préféré à B, B nettement préféré à A)
- Comment évalueriez-vous la dynamique ? (A bien plus équilibré que B, B bien plus équilibré que A)
- Comment évalueriez-vous la réverbération ? (A bien plus naturel que B, B bien plus naturel que A)
- Comment évalueriez-vous la spatialisation ? (A bien plus adapté que B, B bien plus adapté que A)

Ces critères n'ont évidemment pas été choisis au hasard, et ont été déterminés suite à des discussions avec des mixeurs, Etienne Hendrickx ou après les observations effectuées sur les mixages à leur réécoute. En effet, les critères les plus importants qui en sont ressortis sont donc :

- **le timbre**, en particulier la gestion des extrêmes, puisque les aigus sont plus proches grâce au casque et les basses ne sont pas ressenties comme nécessaires en l'absence de LFE ;
- **la dynamique**, qui se retrouve amoindrie lors d'un mixage au casque du probablement au niveau d'écoute assez peu maîtrisé dans ce dispositif, ainsi que par la taille de l'écran, ce qui a pour conséquence de faire ressortir les voix et d'abaisser le niveau des ambiances ;
- **la réverbération**, qui prend une place moins importante dans les mixages faits au casque car l'acoustique reproduite de l'auditorium apporte déjà des réflexions qui ne sont pas en accord avec ce qui est vu par le mixeur dans sa salle, et donc fait l'effet d'une réverbération appliquée au préalable sur l'ensemble du mix ;
- **la spatialisation**, distribuée différemment entre les mixages au casque et en auditorium, où globalement l'image surround est plus large dans les mixages au casque, en particulier sur les enceintes avant bien qu'elle soit normalement resserrée au centre (très probablement un effet de la taille de l'écran).

Toutes les définitions sont bien évidemment rappelées avant chaque session d'écoute. S'il y a des questions, elles sont répondues pour dissiper tout malentendu.

Afin de déterminer si le sujet reconnaît effectivement les extraits entre eux, les mixages « auditorium » et « casque » sont aléatoirement affectés à A ou B, de sorte que le sujet ne savait jamais quel mixage était joué. Le mixage X est proposé après la lecture des deux précédents et demandé à être reconnu par la question « Diriez-vous que X correspond plutôt à B ou plutôt à A ? » et les réponses suivantes : « X est A » ou « X est B », sans possibilité de répondre par « Je ne sais pas ». Cette question est posée lors des deux écoutes.

Les extraits peuvent être écoutés plusieurs fois à la demande du sujet, bien que la plupart du temps les premières écoutes suffisent (Mix A dans son entièreté, puis B, puis X). J'ai d'ailleurs pu observer que des écoutes répétées étaient corrélées avec des erreurs dans la reconnaissance de X : plus le sujet réécoutait d'extraits, plus il se trompait sur la nature de X. Quelques sujets ont déterminé sans faute X sans réécouter une seule fois.

Les extraits mixés allant de 10s à 26s, j'ai pris la décision d'en raccourcir quelques-uns et de faire le tri (donc de manière biaisée) afin de choisir des extraits suffisamment intéressants pour leurs caractéristiques sonores et audiovisuelles mais surtout pour faire baisser le temps de test. Avec cinq extraits allant de 7s à 23s, en

comptant les temps de réponses, réécoutes éventuelles et pause entre les deux écoutes, le test s'étend en tout sur 40 minutes, ce qui est acceptable. Les extraits gardés étaient les extraits 1 et 2 du film de Helio Pu, les extraits 4 et 5 du film d'Arthur Chrisp et l'extrait 7 du film de Louise Giboulot. Ces extraits sont alors renommés extraits 1, 2, 3, 4 et 5.

Comme dit plus haut, c'était à moi de gérer les envois des extraits, mais également leur sélection aléatoire car le dispositif était composé de sessions Pro Tools et de la Studer S5 pour le contrôle. Les mix étaient disposés sous la piste vidéo en trois pistes, la première A, puis B et X. Le passage de l'une à l'autre se faisait par un solo. L'affectation des mixages « auditorium » et « casque » à A ou B était également rendue aléatoire par ma main entre chaque session.

3 – 2. Résultats

Suite à la réception de tous les questionnaires, les données ont été récupérées manuellement dans des tableaux Excel en vue d'une ANOVA selon la norme BS. 1116-3 de l'ITU (Méthodes d'évaluation subjective des dégradations faibles dans les systèmes audio). Les analyses ont été faites par variables dépendantes, qui étaient donc les évaluations pour l'appréciation globale, puis chacune des variables de la deuxième phase, donc le timbre, la dynamique, la réverbération et enfin la spatialisation. Une étude de la corrélation entre les variables a également été menée, notamment pour déceler une justification de l'appréciation par un ou des critères analytiques de la deuxième phase d'écoute.

L'échelle allait de -5 à +5, -5 étant une nette préférence pour le mixage « casque », +5 une nette préférence pour le mixage « auditorium », et 0 aucune préférence.

La p-value permet de déterminer si l'influence d'une variable indépendante sur les réponses des sujets est significative ou non. Les variables indépendantes (ou VI) sont ici les extraits et les mixeurs. Un résultat est considéré comme significatif si sa p-value est inférieure à 5% ou 0.05.

3 – 2. 1. Erreurs

Une erreur correspond à l'incapacité du sujet à reconnaître X. Dans les tableaux ci-contre, les erreurs sont en rouge.

Il est intéressant de remarquer qu'entre les extraits et les écoutes, la répartition des erreurs change drastiquement. Il arrive que des extraits qui présentent un grand nombre d'erreurs à la première écoute n'en présentent aucune à la seconde, et inversement. Certains sujets restent constants dans leurs erreurs entre les deux écoutes, d'autres non.

Pour la première écoute, il y a 24 erreurs pour 150 réponses, soit un taux de réussite de 84%. Il y a eu autant d'erreurs dans la deuxième écoute, le taux est donc le même. Ces taux de réussite élevé (donc d'erreurs bas) suggèrent qu'il y a tout de même des différences significatives entre les deux mixages.

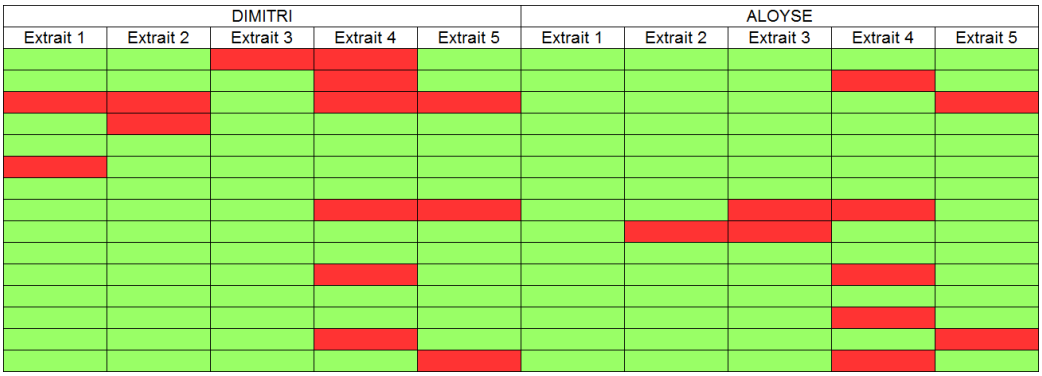


Tableau 1 : Réponses pour la première session.

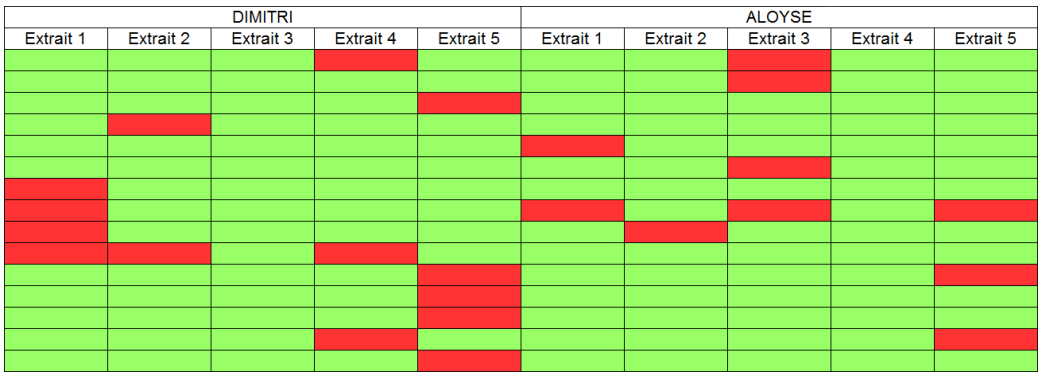


Tableau 2 : Réponses pour la seconde session.

3 – 2. 2. Appréciation

On rappelle que l'appréciation est un jugement sur l'appréciation globale d'un mixage par extrait, soit une évaluation subjective d'un mixage par rapport à l'autre.

Source	*p-value*
Mixeur	0.102
Extrait	**0.001**
Mixeur*Extrait	0.193

Tableau 1 : p-value en fonction des VI pour la variable « appréciation »

On remarque ici que la p-value est inférieure à 0.05 pour la variable « extrait », donc significative (en gras). Cela signifie que l'extrait a une influence sur le jugement de l'appréciation. Un t-test s'impose pour déterminer si les miyennes sont significativement différentes de 0.

Suite au t-test, il y a une moyenne signifcativement différente de 0 pour les extraits 2 et 3, où les mixages au casque sont nettement préférés, les moyennes sont respectivement -1.43 et -1.8. Les extraits 1, 4 et 5 ont des moyennes respectives de 0.5, 0 et -0.37.

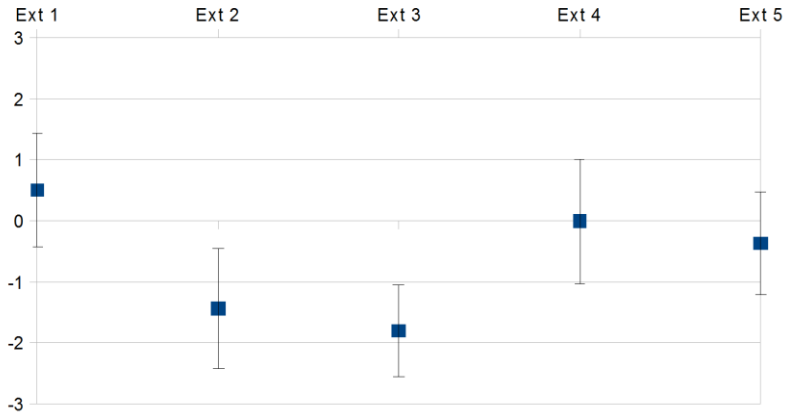


Figure 1. Moyenne des résultats par extrait pour la variable « appréciation » avec intervalle de confiance à 95%

3 – 2. 3. Timbre

On rappelle que le timbre correspond à la gestion fréquentielle du mixage ou son harmonie fréquentielle, pour déterminer si un extrait n'a pas plus de résonances ou de creux qu'un autre.

Source	*p-value*
Mixeur	**0.018**
Extrait	0.120
Mixeur*Extrait	0.406

Tableau 2 : p-value en fonction des VI pour la variable « timbre »

La donnée « mixeur » est ici significative, c'est-à-dire que la qualité du timbre n'a pas été jugée de la même manière par les sujets selon le mixeur. On voit dans la figure 2 que les mixages au casque du mixeur 1 (Dimitri) ont été jugés plus positivement. Le t-test valide en montrant une moyenne significativement différente de 0 pour Dimitri, où ses mixages au casque sont nettement préférés, avec une moyenne de -1.453. Pour Aloyse, la moyenne est à -0.12.

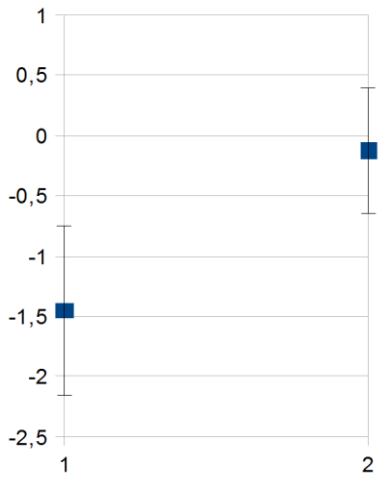


Figure 2. Moyenne des résultats par mixeur pour la variable « timbre » avec intervalle de confiance à 95%

3 – 2. 4. Dynamique

On rappelle que l'évaluation de la dynamique se joue sur l'équilibre des niveaux et la cohérence globale à l'image, c'est-à-dire si un mixage gère mieux sa dynamique sans trop de pics ou de creux, écarts trop grands.

Source	*p-value*
Mixeur	0.268
Extrait	0.396
Mixeur*Extrait	0.760

Tableau 3 : p-value en fonction des VI pour la variable « dynamique »

Ici aucune donnée n'est significative, ce qui veut dire qu'il n'y a pas de différence marquée en fonction de l'extrait ou du mixeur. Le t-test sur l'ensemble des données (mixeurs et extraits confondus) donne une moyenne générale significativement différente de 0, en faveur du mixage au casque, avec une valeur de -0.593.

3 – 2. 5. Réverbération

On rappelle que la réverbération correspond à l'adéquation de la réverbération dans un extrait donné, où le sujet va juger de la pertinence du choix de réverbération ainsi que de sa quantité pour chacun des deux mixages.

Source	*p-value*
Mixeur	0.077
Extrait	0.750
Mixeur*Extrait	0.311

Tableau 4 : p-value en fonction des VI pour la variable « réverbération »

Ces variables ne sont pas significatives, le mixeur ou l'extrait n'ont pas d'influence sur le jugement. Suite au t-test, la préférence est légèrement marquée pour le mixage au casque avec une moyenne globale de -0.5.

3 – 2. 6. Spatialisation

On rappelle que la spatialisation est la donnée d'enveloppement et de clarté spatiale qui doit être également cohérente à l'image.

Source	*p-value*
Mixeur	0.110
Extrait	0.786
Mixeur*Extrait	0.155

Tableau 5 : p-value en fonction des VI pour la variable « spatialisation »

Aucune variable n'est significative, et le t-test n'a pas donné de moyenne significativement différente de 0. Sa valeur est de -0.253. En d'autres termes, il n'y a pas de préférence marquée pour un mixage ou l'autre concernant la variable « spatialisation ».

3 – 2. 7. Corrélation

Une étude de la corrélation a été effectuée entre les diverses variables dépendantes (appréciation, timbre, dynamique, réverbération, spatialisation) pour vérifier si l'appréciation globale était corrélée à l'un des quatre critères d'analyse évalués.

	Appréciation	*Timbre*	*Dynamique*	*Réverb*	*Spatialisation*
Appréciation		0.279	0.264	0.183	0.183
Timbre	0.279		0.359	0.296	0.305
Dynamique	0.264	0.359		0.228	0.321
Réverb	0.183	0.296	0.228		0.461
Spatialisation	0.183	0.305	0.321	0.461	

Tableau 7 : Coefficients de corrélation pour les variables dépendantes

On considère généralement qu'il y a une corrélation forte lorsque les coefficients sont égaux ou supérieurs à 0,8. Or on remarque ici que la valeur la plus élevée est 0,461 entre les variables « Réverbération » et « spatialisation », ce qui semble logique puisque les deux restent sémantiquement liés mais le coefficient reste tout de même faible. En ce qui concerne les relations entre la variable « appréciation » et les autres, le coefficient le plus élevé est 0,279 avec le timbre mais c'est une très petite valeur. On ne peut alors pas déduire de corrélations substantielles entre l'appréciation et l'évaluation de qualité des paramètres analytiques, c'est-à-dire qu'on ne peut pas éventuellement expliquer sur quel(s) paramètre(s) se joue(nt) l'appréciation d'un extrait.

3 – 3. Discussion

On remarque que dans l'ensemble, il y a une préférence pour les mixages qui ont été effectués au casque. Pourtant ce sont des mixages qui ont été faits dans des conditions sub-optimales, en salle de montage, sur petit écran, sans surface de contrôle, avec l'inconfort du casque, etc. Cela pourrait être expliqué par plusieurs choses :

- le mixeur et la mixeuse ont tous deux mixé en premier en auditorium, prenant connaissance des extraits in situ avant de les faire au casque. Comme les séances de mixage au casque ont été plus rapides, on peut penser qu'ils ont été faits en connaissance du résultat in situ.

- il n'y avait qu'une semaine entre chaque séance de mixage, ce qui n'est peut-être pas suffisant pour oublier ce qui a été fait.

- Les voix sont globalement plus claires et il y a moins de réverbération sur les extraits au casque, ce qui pour des scènes courtes hors de leur contexte narratif peut probablement être un avantage.

En regardant plus précisément les moyennes globales, nous pouvons en déduire que les préférences, si elles existent, ne sont pas non plus très marquées et restent proches de l'indifférence (entre -1.5 et -0.5). **On peut en conclure que le transfert est satisfaisant et qu'il n'est pas aberrant, du point de vue du résultat, d'utiliser une reproduction binaurale d'auditorium pour mixer un film.**

Si nous avions eu plus de temps, il aurait été intéressant de proposer à des mixeurs ayant l'habitude de mixer en binaural de participer. Un troisième mixage fait en salle aurait également pu déterminer si les salles de montage ne sont pas déjà suffisantes pour permettre un transfert fiable vers l'auditorium, ou bien un mixage fait au casque mais in situ pour déterminer l'influence visuelle de la salle sur le mixage. Un temps plus long entre chaque séance aurait également pu être préférable. Enfin, un head-tracking des HRTF ainsi qu'une correction de l'EQ du casque à l'écoute auraient pu être implémentés, afin de se rapprocher des « standards » existants.

Conclusion

La situation actuelle ne nous permet pas de deviner une trajectoire certaine pour le mixage au casque en binaural. Malgré la multiplication des offres de plug-ins et hardwares qui permettent ce genre de pratique, il n'est pas assuré pour le moment de les voir se pérenniser dans le domaine de la post-production audiovisuelle. Les apports du binaural ne contrebalancent pas ses désavantages additionnés aux avantages déjà existants dans les habitudes des professionnels du son à l'image.

Cependant, nous avons montré par des tests perceptifs que les mixages effectués au casque en binaural sont jugés cohérents par rapport à ceux faits in situ, en auditorium. Les sujets ont même jugé l'appréciation et les paramètres sonores (timbre, dynamique, réverbération, spatialisation) en moyenne favorablement pour les mixages au casque, bien que ces préférences soient timidement marquées. Nous pouvons en conclure que le transfert est validé, et que la méthode par BRIR est satisfaisante pour reproduire un environnement d'écoute.

Il serait cependant intéressant d'évaluer l'influence des indices visuels dans le travail des monteurs et mixeurs, la taille de l'écran et la lumière semblant être particulièrement importantes dans la perception de l'espace. Un système VR pourrait apporter une solution à ce problème par exemple. Il pourrait également prendre en compte la correction de la réponse en fréquence du casque utilisé ainsi que le tracking de la tête.

Bibliographie

[ABBF20] AGRAWAL, S., SIMON, A., BECH, S., BAERENSTEN, K., FORCHHAMMER, S., Defining Immersion : Literature Review and Implications for Research on Audiovisual Experiences. Dans *Journal of Audio Engineering Society*, vol.68, pp.404-417, juin 2020.

[ADT01] ALGAZI, V. R., DUDA, R. O., THOMPSON, D. M., & AVENDANO, C., The CIPIC HRTF database. Proceedings of the 2001 IEEE Workshop on Applications of Signal Processing to Audio and Electroacoustics. New Paltz, NY, 2001.

[B60] BEKESY, G. von., Experiments in Hearing. New York: McGraw-Hill, 1960.

[B62] BERGEIJK, W. A. van., Variation on a theme of Békésy: A model of binaural interaction. Journal of the Acoustical Society of America, vol 34, pp 1431–1437, 1962.

[B95] BAIBLE, C., L'imparfait du dispositif, Dans *L'Audiophile*, vol. 31, pp.41-51, janvier 1995.

[B97] BLAUERT, J., Spatial Hearing: The Psychophysics of Human Sound Localization. Cambridge, MA: MIT Press, 1997.

[B13] BEZARD, P., *L'insertion d'un outil de spatialisation binaurale dans le flux de post-production et de diffusion sonores*. Ecole Nationale Supérieure Louis-Lumière (Paris, France), 2013.

[BBLMK20] BEST, V., BAUMGARTNER, R., LAVANDIER, M., MAJDAK, P., KOPCO, N., Sound Externalisation : A Review of Recent Research. Dans *Trends in Hearing*, vol.24, pp1-14, 2020.

[BDH13] BRUGHERA, A., DUNAI, L., & HARTMANN, W. M., Human interaural time difference thresholds for sine tones: The high-frequency limit. Journal of the Acoustical Society of America, vol 133, pp 2839–2855, 2013.

[BL16] BIDART, A., LAVANDIER, M., Room-induced cues for the perception of virtual auditory distance with stimuli equalized in level. Acta Acustica united with Acustica, vol 102, pp 159–169, 2016.

[BWA01] BEGAULT, R., WENZEL, M., ANDERSON, R., Direct comparison of the impact of head tracking, reverberation and individualized head-related transfer functions on the spatial perception of a virtual speech source. Journal of the Audio Engineering Society, vol 49, pp 904–916, 2001.

[C02] CHION, M., La scène audio-visuelle. Dans *L'audio-vision* (2e édition). Paris : Nathan, 2002.

[C11] CASTELLANI, F., *Adaptation du mixage cinéma à l'écoute au casque*. Ecole Nationale Supérieure Louis-Lumière (Paris, France), 2011.

[C63] COLEMAN, P., An analysis of cues to auditory depth perception in free space. Psychological Bulletin, vol 60, pp 302–315, 1963.

[CBB18] CUBICK, J., BUCHHOLZ, J. M., BEST, V., LAVANDIER, M., DAU, T., Listening through hearing aids affects spatial perception and speech intelligibility in normal-hearing listeners. Journal of the Acoustical Society of America, vol 144, pp 2896–2905, 2018.

[CBR14] CONETTA, R., BROOKES, T., RUMSEY, F., ZIELINSKI, S., DEWHIRST, M., JACKSON, P., BECH, S., MEARES, D., & GEORGE, S., Spatial audio quality perception (Part 2): A linear regression model. Journal of Audio Engineering Society, vol 62, pp 847–860, 2014.

[CCA13] CALLAN, A., CALLAN, D., ANDO, H., Neural correlates of sound externalization. NeuroImage, vol 66, pp 22–27, 2013.

[CSB13] CATIC, J., SANTURETTE, S., BUCHHOLZ, J. M., GRAN, F., DAU, T., The effect of interaural-level-difference fluctuations on the externalization of sound. Journal of the Acoustical Society of America, vol 134, pp 1232–1241, 2013.

[G15] GEDEMER, L., Subjective Listening Tests for Preferred Room Response in Cinemas – Part 1 : System and Test Descriptions. Présenté à la 139th AES Convention, New-York, USA, octobre 2015.

[G16] GEDEMER, L., Subjective Listening Tests for Preferred Room Response in Cinemas – Part 2 : Preference Test Results. Présenté à la 140th AES Convention, Paris, France, 4-7 juin 2016.

[GCSD16] GIL-CARVAJAL, J-C., CUBICK, J., SANTURETTE, S., and DAU, T., Spatial Hearing With Incongruent Visual or Auditory Room Cues. Dans *Nature Sci. Rep*., vol 6, p. 37342, 2016.

[GG73] GARDNER, B., GARDNER, S., Problem of localization in the median plane: Effect of pinnae cavity occlusion. Journal of the Acoustical Society of America, vol 53, pp 400–408, 1973.

[GM94] GARDNER, B., MARTIN, K., HRTF Measurements of a KEMAR Dummy-head Microphone. Cambridge : Massachusetts Institute of Technology, 1994.

[H17] HENDRICKX, E., STITT, P., MESSONIER, J-C., LYZWA, J-M., KATZ, B., de BOISHERAUD, C., Influence of head tracking on the externalization of speech stimuli

for non-individualized binaural synthesis. Dans *Journal of Audio Engineering Society*, vol 141, pp 2011-2023, mars 2017.

[H90] HOLMAN, T., Surround Sound Systems Used With Pictures in Cinemas and Homes. Présenté à la 8th AES International Conference, mai 1990.

[HGF03] HUNTER, M., GRIFFITHS, T., FARROW, T., ZHENG, Y., WILKINSON, I., HEGDE, N., WOODS, W., SPENCE, S., WOODRUFF, P., A neural basis for the perception of voices in external auditory space. Brain, vol 126(Pt 1), pp 161–169, 2003.

[HV03] HOFMAN, P., & VAN OPSTAL, A., Binaural weighting of pinna cues in human sound localization. Experimental Brain Research, vol 148, pp 458–470, 2003.

[HW96] HARTMANN, W., WITTENBERG, A., On the externalization of sound images. Journal of the Acoustical Society of America, vol 99, pp 3678–3688, 1996.

[HWD17] HASSAGER, H. G., WIINBERG, A., DAU, T., Effects of hearing-aid dynamic range compression on spatial perception in a reverberant environment. Journal of the Acoustical Society of America, vol 141, pp 2556–2568, 2017.

[J48] JEFFRESS, L. A., A place theory of sound localization. Journal of Comparative and Physiological Psychology, vol 41, pp 35–39, 1948.

[KHB12] KOPCO, N., HUANG, S., BELLIVEAU, J., RAIJ, T., TENGSHE, C., AHVENINEN, J., Neuronal representations of distance in human auditory cortex. Proceedings of the National Academy of Sciences of the United States of America, vol 109, pp 11019–11024, 2012.

[L89] LETOWSKI, T., Sound quality assessment: Cardinal concepts. Proceedings of the 87th Audio Engineering Society Convention. Hamburg, Germany, 1989.

[LVM02] LARSSON, P., VASTFJALL, D., KLEINER, M., Auditory- ¨ Visual Interaction in Real and Virtual Rooms. Présenté à *Proceedings of the Forum Acusticum, 3rd EAA European Congress on Acoustics*, 2002.

[M58] MILLS, A., On the minimum audible angle. Journal of the Acoustical Society of America, vol 30, pp 237, 1958.

[M72] MILLS, A., Auditory localization (Binaural acoustic feld sampling, head movement and echo effect in auditory localization of sound sources position, distance and orientation). *Foundations of Modern Auditory Theory*, vol 2, pp 303–348, 1972.

[M92] MOLLER, H., Fundamentals of binaural technology. Dans *Applied Acoustic*s, vol 36, p171-218, 1992.

[MFT99] MACKENSEN, P., FELDERHOFF, U., THEILE, G., Binaural Room Scanning – A new Tool for Acoustic and Psychoacoustic Research. Dans *Journal of the Acoustical Society of America*, vol.105, 1999.

[MK10] MAGEZI, D. A., KRUMBHOLZ, K., Evidence for opponent-channel coding of interaural time differences in human auditory cortex. *Journal of Neurophysiology*, vol 104, pp 1997–2007, 2010.

[ML10] MEDDIS, R., & LOPEZ-POVEDA, E. A., Auditory periphery: From pinna to auditory nerve. Dans Meddis et al. (Eds.), Computational Models of the Auditory System, New York: Springer, 2010.

[MOC01] MINNAAR, P., OLESEN, S., CHRISTENSEN, F., MØLLER H., Localization with Binaural Recordings from Artificial and Human Heads, *Journal of the Acoustical Society of America*, vol. 49, pp. 323–336, mai 2001.

[NSMH03] NOISTERNIG, M., SONTACCHI, A., MUSIL, T., HOLDRICH, R., A 3D Ambisonic Based Binaural Sound Reproduction System. Présenté à la AES 24th International Conference : Multichannel Audio, The New Reality, Alberta, Canada, juin 2003.

[ON20] ORAMUS, T., NEUBAUER, P., Comparison of Perception of Spatial Localization Between Channel and Object Based Audio. Présenté à la 148th AES convention, 2-5 juin 2020.

[OP84] OLDFELD, R., PARKER, P., Acuity of sound localization: A topography of auditory space: II: Pinna cues absent. Perception, vol 13, pp 601–617, 1984.

[OWM07] OLIVE, S., WELTI, T., MARTENS, W., "Listener Loudspeaker Preference Ratings Obtained In Situ Match Those Obtained via a Binaural Room Scanning Measurement and Playback System". Présentée à la 122nd AES Convention. Vienne, Autriche, 2007.

[P74] PLENGE, G., On the differences between localization and lateralization. Journal of the Acoustical Society of America, vol 56, pp 944–951, 1974.

[PNHR87] PATTERSON, R., NIMMO-SMITH, I., HOLDSWORTH, J., RICE, P., An efficient auditory filterbank based on the gammatone function. Présenté au Meeting of the IOC Speech Group on Auditory Modelling at RSRE, 1987.

[PS90] PERROTT, D., SABERI, K., Minimum audible angle thresholds for sources varying in both elevation and azimuth. Journal of the Acoustical Society of America, vol 87, pp 1728–1731, 1990.

[RG17] ROGINSKA, A. GELUSO, P., *Immersive Sound : The art and science of binaural and multi-channel audio.* Royaume-Uni : Taylor & Francis, 2017.

[RLR08] RICHARDSON, G., LUKASHKIN, A., RUSSELL, I., The tectorial membrane: One slice of a complex cochlear sandwich. Current Opinion in Otolaryngology & Head and Neck Surgery, 2008.

[RZKB05] RUMSEY, F., ZIELINSKI, S., KASSIER, R., BECH, S., Relationships between experienced listener ratings of multichannel audio quality and naïve listener preferences. Journal of Acoustical Society of America, vol 117, pp 3832–3840, 2005.

[SDC08] SHUB, D., DURLACH, N., COLBURN, H., Monaural level discrimination under dichotic conditions. Journal of the Acoustical Society of America, vol 123, pp 4421–4433, 2008.

[SEW19] SCHUTTE, M., EWERT, S., WIEGREBE, L., "The Percept of Reverberation Is Not Affected by Visual Room Impression in Virtual Environments,. Dans *Journal of the Acoustic Society of America*, vol 145, no. 3, pp. EL229–EL235, 2019.

[SHEP20] SALMON, F., HENDRICKX, E., EPAIN, N., PAQUIER, M., The Influence of Vision on Perceived Differences Between Sound Spaces. Dans *Journal of the Audio Engineering Society*, vol.68, pp.522-531, juillet-août 2020.

[SK10] SCHONSTEIN, D., KATZ, B., Sélection de HRTF dans une Base de Données en Utilisant des Paramètres Morphologiques pour la Synthèse Binaurale. Présenté au 10ème Congrès Français d'Acoustique, avril 2010, Lyon, France.

[S74] SHAW, E. A. G., The external ear. In W. D. Keidel & W. D. Neff (eds.), Handbook of Sensory Physiology, Vol. 5/1, Auditory System. New York: Springer Verlag, 1974.

[SSCK10] SMYTH, S., SMYTH, M., CHEUNG, S., KRAMER, L., A Virtual Acoustic Film Dubbing Stage. Présenté à la 40th AES Conference à Tokyo, Japon, 2010.

[STYM10] SALMINEN, N., TIITINEN, H., YRTTIAHO, S., MAY, P., The neural code for interaural time difference in human auditory cortex. Journal of the Acoustical Society of America, vol 127, pp 60–65, 2010.

[TR67] THURLOW, W. R., & RUNGE, P. S. (1967). Effect of induced head movements on localization of direction of sounds. Journal of the Acoustical Society of America, vol 42, pp 480–488, 1967.

[VLV04] VALJAMAE, A., LARSON, P., VASTFJALL, D., KLEINER, M., Auditory Pressure, Individualized Head-Related Transfer Function, and Illusory Ego-Motion in Virtual Environments, Proceedings of the Seventh Annual Workshop in Presence, Spain, 2004.

[W38] WOODWORTH, R., Experimental Psychology. New York: Holt, 1938.

[W39] WALLACH, H., On sound localization. Journal of the Acoustical Society of America, vol 10, pp 270–274, 1939.

[WAK93] WENZEL, E., ARRUDA, M., KISTLER, D., WIGHTMAN, F., Localization Using Nonindividualized HeadRelated Transfer Functions, *Journal of the Acoustical Society of America*, vol 94, pp. 111–123, 1993.

[WK99] WIGHTMAN, L., KISTLER, D., Resolution of front-back ambiguity in spatial hearing by listener and source movement. *Journal of the Acoustical Society of America*, vol 105, pp 2841–2853, 1999.

[WRS11] WIERSTORF, H., GEIER, M., RAAKE, A., & SPORS, S., A free database of head-related impulse response measurements in the horizontal plane with multiple distances. Proceedings of the 130th AES Convention, eBrief. London, UK, 2011.

[WS54] WOODWORTH, R., SCHLOSBERG, H., Experimental Psychology. New York: Holt, 1954.

[XZY13] XIE, B., ZHONG, X., YU, G., GUAN, S., RAO, D., LIANG, Z., & ZHANG, C., Report on Research Projects on Head-related transfer functions and virtual auditory displays in China. *Journal of Audio Engineering Society*, vol 61, pp 314–326, 2013.

[YD97] YOST, W. A., & DYE, R. H., Fundamentals of directional hearing. Seminars in Hearing. New York: Thieme Medical Publishers, 1997.

Annexes

Première écoute

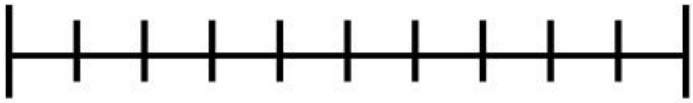
Extrait 1 :

- Diriez-vous que X correspond plutôt à B ou plutôt à A ? (entourez la réponse)

X est A **X est B**

- Comment évalueriez-vous votre préférence entre A et B ? (entourez la gradation)

A nettement préféré à B **B nettement préféré à A**



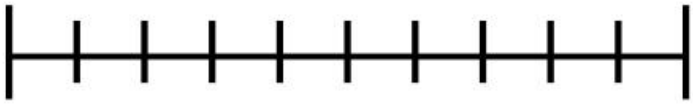
Extrait 2 :

- Diriez-vous que X correspond plutôt à B ou plutôt à A ? (entourez la réponse)

X est A **X est B**

- Comment évalueriez-vous votre préférence entre A et B ? (entourez la gradation)

A nettement préféré à B **B nettement préféré à A**



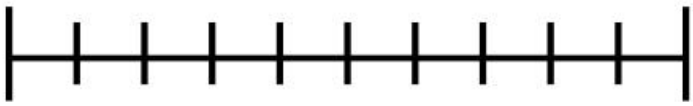
Extrait 3 :

- Diriez-vous que X correspond plutôt à B ou plutôt à A ? (entourez la réponse)

X est A **X est B**

- Comment évalueriez-vous votre préférence entre A et B ? (entourez la gradation)

A nettement préféré à B **B nettement préféré à A**



Annexe 1 : Page 1 du questionnaire utilisé dans le test perceptif

Deuxième écoute

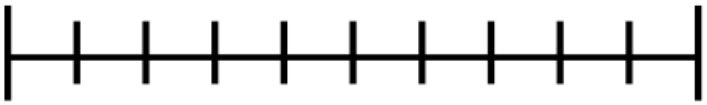
Extrait 1 :

- Diriez-vous que X correspond plutôt à B ou plutôt à A ? (entourez la réponse)

X est A **X est B**

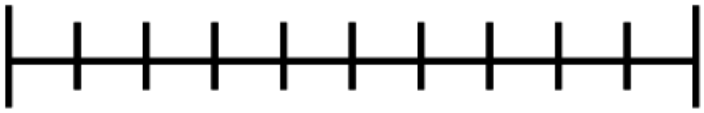
- Comment évalueriez-vous la qualité du timbre entre A et B ? (entourez la gradation)

A nettement préféré à B **B nettement préféré à A**



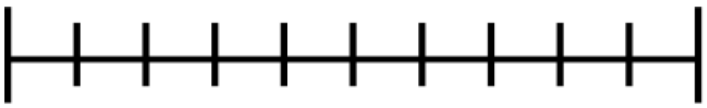
- Comment évalueriez-vous la dynamique ? (entourez la gradation)

A bien plus équilibré que B **B bien plus équilibré que A**



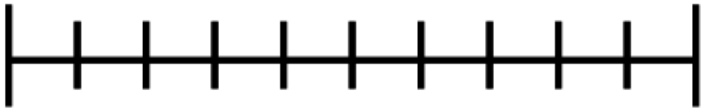
- Comment évalueriez-vous la réverbération ? (entourez la gradation)

A bien plus naturel que B **B bien plus naturel que A**



- Comment évalueriez-vous la spatialisation ? (entourez la gradation)

A bien plus adapté que B **B bien plus adapté que A**



Annexe 2 : Page 5 du questionnaire utilisé dans le test perceptif