Saillance auditive : de la caractérisation psychoacoustique à la perception de l'environnement sonore

Baptiste Bouvier

Thèse doctorale de Sorbonne Université

École Doctorale SMAER

Saillance auditive

De la caractérisation psychoacoustique à la perception de l'environnement sonore

Auteur
Baptiste **Bouvier**

Directeurs
Nicolas **Misdariis** – Dir. de Recherche, STMS Lab, IRCAM
Catherine **Marquis-Favre** – Dir. Recherche, ENTPE

Encadrant
Patrick **Susini** – Dir. de Recherche, STMS Lab, IRCAM

Rapporteurs
Nicolas **Grimault** – Dir. de Recherche, CRNL, CNRS, *président du jury*
Arnaud **Can** – Dir. de Recherche, UMRAE

Examinateurs
Mounya **Elhilali** – Professor, Johns Hopkins University
François **Ollivier** – Maître de Conférences, Sorbonne Université
Sabine **Meunier** – Chargée de Recherche, HdR, LMA, CNRS

31 janvier 2024

“L’attention, en tout, c’est ce qui nous sauve.”

Bossuet

Résumé

Au coeur des mécanismes cognitifs mis en jeu dans la perception de notre environnement, la saillance traduit la capacité de certaines informations à capturer notre attention indépendamment de notre volonté. Si la saillance auditive semble façonner notre perception de l'environnement sonore, son étude relève d'enjeux variés : pollution sonore dans l'environnement urbain, conception de sources audibles pour la transmission d'informations utiles se trouvent ainsi en miroir autour de cette notion.
Dans ces travaux, nous nous demandons comment les propriétés des sources sonores sont susceptibles de moduler leur saillance, et comment leur saillance est susceptible d'affecter notre perception de la scène sonore dans sa globalité. Entre autres, nous souhaitons observer comment elle affecte notre perception de l'agrément dans un paysage sonore.
Nous révélons ainsi en premier lieu la composante *stimulus-driven* de l'attention, plus précisément la modulation d'un phénomène de capture attentionnelle par des attributs du timbre dans des séquences sonores contrôlées. Puis, nous montrons comment la présence de sons saillants affecte la perception de séquences sonores plus complexes en y révélant l'effet de la saillance à un niveau local sur la primauté du traitement holistique habituellement observé. Enfin, nous étudions le lien entre saillance et agrément sonore dans des scènes environnementales variées, établissant la saillance comme indicateur essentiel de la perception et l'appréciation de l'environnement sonore.
À la lumière des résultats obtenus, il apparaît que notre environnement sonore peut s'imposer à nous, par le biais des éléments saillants qui en émergent et en façonnent notre perception et notre appréciation.

Abstract

At the core of the cognitive processes involved in perceiving the environment, salience refers to the ability of certain information to capture our attention independently of our will. Auditory salience appears to shape our perception of the soundscape and its exploration involves various challenges : noise pollution in the urban environment and designing audible sources for the transmission of useful information are thus mirrored around this notion. In this study, we examine how the features of sound sources are likely to modulate their salience, and how their salience is likely to affect our perception of the surrounding scene. Specifically, we investigate how the salience of sound sources influences our perceptual experience of pleasantness in a soundscape.

We first reveal the *stimulus-driven* component of attention, more precisely the modulation of an attentional capture phenomenon by timbre attributes in controlled sound sequences. We then demonstrate the influence of salient sounds on the perception of more complex sound sequences, unveiling the effect of salience at a local level on the typically observed prioritization of holistic processing. Furthermore, we explore the link between temporal salience and pleasantness in a variety of soundscapes, confirming salience as a crucial indicator for the assessment and the perception of sonic environments.

In the light of our findings, we note that soundscapes are able to impose themself on us through the salient elements that emerge and shape our perception and appreciation of it.

Remerciements

L'exposé, dans ce manuscript, d'une partie des travaux réalisés ces trois dernières années ne saurait résumer la richesse des expériences qui ont marqué cette période. Je tiens donc, avant de solliciter l'attention du lecteur que j'essayerai tant bien que mal de maintenir sur le contenu de cette thèse, à remercier les personnes qui m'ont accompagné.

En premier lieu, j'aimerais remercier mon équipe d'encadrants, qui m'a fait confiance dans la conception et la réalisation de ce projet doctoral. Je remercie Nicolas Misdariis pour la bienveillance systématique dont il a fait preuve dans l'encadrement de ce projet, comme de tous ceux de l'équipe Perception et Design Sonores, ainsi que pour la confiance qu'il a placée dans ce projet dès ma première sollicitation, début 2019, bien avant le début de cette thèse. Je remercie Catherine Marquis-Favre d'avoir su porter avec confiance ce projet, depuis ma première sollicitation également, et d'être parvenue à l'encadrer à distance avec une rigueur que j'espère avoir acquise aujourd'hui. Je remercie enfin Patrick Susini pour ses conseils, son soutien, et les stimulations dont il m'a régulièrement fait part. Merci à vous pour la liberté que vous avez su m'accorder et l'adaptabilité dont vous avez fait preuve durant cette expérience d'encadrement. J'espère pouvoir collaborer avec vous au-delà de cette thèse. Je remercie également le corps des Ingénieurs des Ponts, des Eaux et des Forêts, et plus précisément la Commission de Formation Doctorale, d'avoir financé ce projet doctoral, en m'accordant une grande confiance et une grande liberté.

Je remercie Emmanuel Ponsot pour sa participation aux travaux du chapitre 3, ainsi que pour les échanges souvent emballés autour de paradigmes expérimentaux divers et variés. Merci également à Pablo Arias pour le soutien durant toute la première partie de cette thèse, et notamment la découverte du monde expérimental et de ses ressorts. À ce propos, je ne pourrais omettre

de remercier toute l'équipe de l'INSEAD, entre autres Huong, Germain et Sébastien, grâce à qui j'ai pu mener un certain nombre d'expériences dans les meilleurs conditions (presque 200 participants pour autant d'heures d'expériences chez vous !).

Je souhaite également remercier tous les doctorants passés par l'équipe Perception et Design Sonores au cours de ces années. Merci donc à Nadia pour les traditions du nord dont un ch'ti avait bien besoin, Valérian pour un CFA 2020 riche en émotions ainsi que de belles discussions et perspectives scientifiques, Victor pour l'animation de l'équipe PDS et les leçons musicales, Claire pour l'acceptation de séances de beat-box frénétiques. Merci également à ceux ayant fait symboliquement partie de l'équipe, notamment Constance pour les parties de uno endiablées, et Yann pour la passion, le temps, et les belles collaborations pour le plaisir. Je remercie également Romain, Matthieu et Emma d'avoir permis de croiser nos domaines respectifs et de donner lieu à des collaborations originales. Merci enfin à Clara que j'ai eu la chance de pouvoir encadrer en stage.

Enfin, à ceux qui ont soutenu ce projet indirectement mais qui m'ont soutenu plus personnellement, un grand merci également. Merci à mes parents de m'avoir soutenu dans l'émergence de cette idée de thèse dans le monde sonore et permis de transformer une intuition fugace en éventualité. Merci à Amélie pour le précieux soutien dans la dernière ligne droite ; l'idée de la prolonger ensemble m'emplit de joie. Enfin, merci à Florian, Matthias, LV, Alban, Alfred, Bastien, Nicolas, Sarah, Romane, Guillaume et tous ceux que je ne peux pas nommer ici de manière exhaustive, d'avoir été tant présents pour partager et accompagner les péripéties traversées durant ces années.

Table des matières

INTRODUCTION

Vibrer.

Pour tout être vivant sur cette Terre, l'air qui insuffle la vie est aussi celui qui trahit à chaque instant. Car toute activité met en branle le monde autour de soi. Chaque organisme irradie son environnement de vibrations au moindre mouvement, et chacun se trouve ainsi immergé sous les flots des perturbations aériennes issues de son environnement. L'évolution du vivant a ainsi doté nombre de ses protagonistes d'organes de plus en plus sophistiqués destinés à capter ces vibrations. Chez l'Homme, cet organe atteint des prouesses exceptionnelles : nos oreilles sont capables de détecter des variations de pression 10 milliards de fois plus faibles que la pression atmosphérique, et captent tout mouvement de l'air qui oscille entre 20 et 20000 fois par seconde. Les performances de notre système auditif dépassent celles de tous les autres sens : nous percevons des sons avec une intensité balayant 12 ordres de grandeur (de 10^{-12} à 1 W/m^2), et ce dans toutes les directions. C'est pourquoi on qualifie parfois l'ouïe de sentinelle des sens.

Entendre.

Or, ce miracle de l'évolution en cache un autre, peut-être plus grand encore. Derrière nos oreilles, sous la partie émergée de l'iceberg, notre cerveau accomplit à chaque instant une oeuvre colossale que l'expérience sonore consciente ne laisse pas transparaître. Car notre système sensoriel ultra-performant ne se lasse pas de réagir dans un environnement qui n'est jamais silencieux, et seul un ensemble élaboré de traitements cognitifs nous permet de surmonter ce débordement permanent d'informations. À tout instant, notre cerveau filtre

ainsi la plupart des sollicitations qui nous parviennent pour que nous puissions nous concentrer sur celles, peu nombreuses en comparaison, qui nous semblent pertinentes. Plus ingénieux encore, notre système cognitif a appris que certaines informations que nous n'avons pas prévu d'entendre peuvent parfois se révéler cruciales. Il se laisse donc la possibilité de permettre à des sons inattendus de s'imposer à nous indépendamment de nos intentions. Nous ne sommes donc pas pleinement maîtres de notre perception. Pendant des millions d'années, les mécanismes permettant d'établir l'équilibre subtil entre filtrage et ouverture aux sollicitations extérieures ont progressé.

S'immerger.

Or, l'environnement dans lequel nous évoluons aujourd'hui n'a plus grand chose à voir avec celui dans lequel nous avons appris à entendre. Les mécanismes cognitifs réglés minutieusement dans nos environnements passés ne sont plus ajustés au monde dans lequel nos oreilles sont actuellement immergées. Ainsi, alors que nous sommes conçus pour permettre à certaines sollicitations sonores de s'imposer à nous, nous avons au fil du temps chargé, surchargé, saturé l'environnement sonore, dont les hurlements nous harcèlent de plus en plus. Dans ce nouvel univers acoustique, l'oreille humaine est assaillie par des bruits toujours plus nombreux et puissants. Réglé naturellement pour transmettre la détection d'infimes variations et de subtiles signatures acoustiques, notre système perceptif et cognitif permet aujourd'hui à notre environnement sonore de nous submerger.

Comprendre.

Comprendre le rapport de l'homme à son environnement sonore demande ainsi de percer les mystères des subtils mécanismes cognitifs qu'il déploie depuis ses premiers pas : **comment certains sons parviennent-ils à s'imposer à nous ? Comment leur présence affecte notre perception et notre appréciation de l'environnement sonore ?**

Ces questions sont si vastes qu'il n'est aujourd'hui pas possible d'y répondre précisément. Pourtant, notre environnement évolue à une vitesse extraordinaire au regard des temps biologiques qui ont façonné notre système sensoriel. Nous nous retrouvons ainsi confrontés à des problèmes dans notre relation à l'environnement auxquels nous ne sommes pas capables de proposer de solution immédiate.

Ainsi, ma mission réalisée en 2020 au sein de la mission bruit et agents physiques de la Direction Générale de la Prévention des Risques (DGPR) du Ministère de la Transition Écologique et de la Cohésion des Territoires (MTECT) se soldait par la constatation suivante : face au besoin grandissant d'une action des pouvoirs publics sur le sujet de l'exposition au bruit, le manque de connaissance ne permet pas d'agir de manière pertinente. De fait, la loi d'orientation des mobilités (LOM) [1] consacrait quelques mois plus tôt la notion de pollution sonore, prévoyait que soient précisées les modalités d'évaluation des pics de bruit causés par les transports ferroviaires, et affirmait le droit de chacun à vivre dans un "environnement sonore sain". Pourtant, il n'existait pas de littérature suffisamment riche et de consensus scientifique sur le sujet pour que l'administration puisse réglementer sur ces sujets, entre autres sur le cas des expositions sonores du type de celles causées par des infrastructures ferroviaires à l'origine de pics de bruit. Nous concluions donc cette mission en recommandant, auprès de la DGPR et du Conseil National du Bruit, de mener des travaux de recherche permettant de mieux appréhender les situations d'exposition aux bruits saillants.

Ce fut chose faite lorsque le CEIGIPEF (Centre Interministériel de Gestion des Ingénieurs des Ponts, des Eaux et des Forêts) finança ce projet de thèse proposé à mon initiative, en collaboration entre l'IRCAM, institut de recherche en acoustique, et l'ENTPE, école d'ingénieurs de l'aménagement durable des territoires sous tutelle du MTECT. Le contexte dans lequel ce projet de thèse fut conçu et les raisons qui le motivèrent furent exposés et discutés en amont de son financement et sont rappelés ici.

1. Loi n° 2019-1428 du 24 décembre 2019

Enjeux sociétaux

En 2011, l'Organisation Mondiale de la Santé (OMS, 2011) indiquait qu'entre 1 et 1,6 millions d'années en bonne santé seraient perdues (en anglais : "disability-adjusted life-years", ou DALYs) chaque année en Europe occidentale sous l'effet du bruit. Le bruit des transports y serait responsable d'au moins 10 000 cas de mortalité prématurée et de 43 000 hospitalisations par an. D'après l'étude, parmi les facteurs de risque environnemental en Europe, la surexposition aux bruits serait ainsi la seconde cause de morbidité derrière la pollution atmosphérique.

Selon un sondage de l'Institut Français d'Opinion Publique (IFOP, 2014) commandé en 2014 par le Ministère de l'Écologie, du Développement Durable et de l'Énergie (MEDDE), 82% des Français indiquent se préoccuper des nuisances sonores. Un sondage réalisé pour la Journée Nationale de l'Audition (JNA, 2016) ajoute que la quasi-totalité des personnes sondées (94%) pense que le bruit a des effets directs sur la santé et que plus de 9 Français sur 10 se disent exposés chaque jour à un bruit qu'ils jugent excessif. Enfin, plus de 8 personnes sur 10 attendent une impulsion des pouvoirs publics afin qu'ils prennent mieux en compte l'impact du bruit dans leur vie quotidienne et sur leur santé. Par ailleurs, le bruit est également considéré comme la première préoccupation relative à la qualité de vie, et comme facteur de stress dans l'espace urbain [2].

Le Ministère de la Transition Écologique et de la Cohésion des Territoires (MTECT) est de fait doté d'un organisme, le Conseil National du Bruit (CNB), qui, dans une étude menée avec l'Agence de l'Environnement et de la Maîtrise de l'Énergie en 2016, puis révisée en 2021, (ADEME, 2021), évaluait le coût du bruit dans notre société à 147 milliards d'euros par an. Ce coût social correspond aux coûts engendrés par différents effets du bruit : dépenses pour traiter les impacts sanitaires, pertes de productivité, dépréciation immobilière, etc. Les deux tiers de ce coût sont liés au bruit des transports, qui causent

2. Enquête de l'INSEE-SOeS, citée par le Commissariat Général au Développement Durable dans leur bulletin de mars 2014

entre autres troubles du sommeil, de l'apprentissage, gêne, perte de productivité, ou maladies cardio-vasculaires.

L'Union Européenne a d'abord proposé de baser son action pour la lutte contre les nuisances sonores sur l'utilisation du L_{DEN} (un indice moyennant le niveau sonore sur 24 heures avec des pondérations plus importantes le soir et la nuit), notamment via la directive 2002/49/CE de 2002 relative à l'évaluation et à la gestion du bruit dans l'environnement. Cet indice a l'avantage d'être facile à calculer à partir du niveau équivalent pondéré A $L_{A,eq}$ (pondération fréquentielle du niveau sonore qui prend en compte la courbe de sensibilité en fréquence de l'oreille humaine). Cependant, Berglund (1998) ou Lercher (1998) ont montré qu'une faible partie de la gêne, moins d'un tiers, serait explicable par le niveau sonore $L_{A,eq}$. Il est donc nécessaire de trouver de meilleurs indicateurs permettant de caractériser l'effet de la pollution sonore que subissent les individus.

Le Conseil Général de l'Environnement et du Développement Durable (CGEDD, 2017) a ainsi noté que « le bruit dans l'environnement se caractérise [...] par sa très grande variabilité : le bruit de fond constitué par la circulation varie dans la journée, une mobylette qui passe dans un environnement calme apparaît comme un bruit émergent dans un bruit de fond général, et des bruits de même intensité peuvent être de qualité très différente. La répétition de bruits émergents, même d'intensité modérée, peut s'avérer très pénible à supporter. Il est donc particulièrement difficile de définir un indicateur qui reflète cette complexité, en particulier celle des bruits émergents ». De tels enjeux se sont reflétés dans la réglementation récente : des articles concernant la nécessaire prise en considération des pics de bruit et l'inscription dans le code de l'environnement du droit de vivre dans un environnement sonore sain ont en effet été intégrés dans la Loi d'Orientation des Mobilités en 2019.

La réglementation se heurte encore pour le moment à un manque de connaissances sur le sujet. L'ANSES (2013) (l'Agence Nationale de Sécurité Sanitaire, de l'alimentation, de l'environnement et du travail) soulignait

ainsi, dans un rapport d'expertise, qu'il n'est « pas possible de déterminer des indicateurs opérationnels (...) en raison d'une part des lacunes dans les connaissances actuelles et d'autre part de la complexité des interactions entre les divers paramètres physiques, physiologiques, humains et cognitifs impliqués dans les relations bruit-santé ». Bien que de premières tentatives de prise en compte des pics de bruit émergent, les outils de cartographie ou de mesure de l'exposition sonore sont essentiellement fondés sur des niveaux d'énergie acoustique moyens. Des travaux de recherche sont nécessaires pour mieux comprendre la notion de saillance auditive, responsable de l'émergence de certains bruits et à l'interface entre paramètres physiques, perceptifs et cognitifs (cf. chapitre 1), et qui semble être une composante essentielle de notre perception des bruits environnementaux.

Cadre d'étude

Ainsi, les motivations originelles de ces travaux sont issues de préoccupations concernant l'environnement sonore. Ce dernier, comme dans bien d'autres domaines à l'interface entre l'homme et son environnement, semble en effet affecter de manière irrépressible les individus qui y évoluent. Bien que des moyens techniques nous permettent de s'abstraire des sollicitations issues de l'environnement (casques, écouteurs, double et triple vitrages, protections acoustiques, etc), il semble en effet que les assauts d'informations sonores s'imposent à nous de plus en plus. Si l'on souhaite pouvoir comprendre notre rapport à cet environnement sonore et agir sur son contenu pour préserver notre bien-être, il semble ainsi nécessaire de s'intéresser aux facteurs susceptibles de rendre des sources sonores saillantes dans leur environnement, et aux effets de cette saillance sur la perception et l'appréciation de cet environnement.

Ce sujet peut également être appréhendé en miroir de ce point de vue. L'environnement anthropique étant de plus en plus riche en sollicitations sensorielles, une information pertinente doit se démarquer dans un paysage

sonore parfois saturé pour être perçue et traitée par un individu. Il est alors intéressant de se demander quels mécanismes peuvent permettre à cette information sonore d'émerger, et comment cette information peut affecter la perception de l'environnement sonore.

L'étude de la saillance auditive nécessite avant toute chose de s'interroger sur la définition de ce concept. Car si le terme "saillant" est utilisé dans le langage courant, sa définition sur le plan scientifique comprend certaines subtilités. Nous verrons en partie 1.2 que la définition adoptée dans cette thèse est *la capacité d'un son à capturer l'attention d'un sujet*. Plus précisément, plus un son est saillant, plus il est susceptible de capturer l'attention, et ce, indépendamment des intentions de l'individu. Le concept d'attention mentionné ici mérite lui aussi une définition qui sera précisée et discutée en partie 1.1. Quoi qu'il en soit, ces notions touchent au concept de fonction cognitive, qui permettent à l'humain de percevoir son environnement et de traiter les informations qui en proviennent pour adopter un comportement adéquat. Il nous faut ainsi, pour étudier et comprendre ces concepts, passer du domaine de l'acoustique à celui de la psychologie cognitive. Ce n'est qu'à la lumière des méthodes de cette discipline que nous pourrons tirer des résultats nous permettant de mieux comprendre les aspects déterminants de la saillance auditive qui nous intéressent.

C'est pourquoi le propos est introduit, au fil des parties 1.1 et 1.2, en explicitant d'abord le concept d'attention comme fonction cognitive majeure, puis de saillance comme composante de cette dernière. Notre intérêt se portant sur l'influence d'éléments saillants dans des scènes sonores environnementales, nous introduisons en partie 1.3 certains concepts propres à l'analyse de scènes auditives complexes, puis à la perception de scènes sonores environnementales en partie 1.4. Enfin, nous résumons en partie 1.5 les questionnements liés à nos préoccupations et émergeant de la littérature, et nous présentons la structure de la thèse mise en oeuvre pour y répondre.

Chapitre 1

État de l'art

"On fait la science avec des faits, comme on fait une maison avec des pierres ; mais une accumulation de faits n'est pas plus une science qu'un tas de pierres n'est une maison."

Henri Poincaré

1.1 L'attention

L'attention est un terme courant dans le langage quotidien. On s'exclame "attention !" pour alerter d'un danger, on prête son attention à quelque chose dont on veut tenir compte, certaines personnes souffrent de troubles de l'attention. On peut encore capter l'attention d'un auditoire, ou faire attention à ses dépenses. Le dictionnaire du Petit Robert définit l'attention comme la "concentration de l'activité mentale sur un objet". Ses différents usages suggèrent qu'il s'agit d'un concept que l'on peut orienter, doser, diviser ou déclencher.

L'attention est un concept particulièrement étudié en psychologie cognitive. Cette discipline est une branche de la psychologie qui appartient aux sciences cognitives. Elle s'intéresse aux grandes fonctions psychologiques de l'humain qui lui permettent d'analyser les informations issues de son environnement et de s'en faire une représentation mentale interne pour pouvoir comprendre, décider et agir. Ainsi, l'attention y est étudiée aux côtés de la mémoire, du langage, du raisonnement ou des émotions. De fait, celle-ci est mobilisée dans la plupart des situations que nous rencontrons et agit à l'interface des autres fonctions cognitives.

Si l'attention fut un sujet d'importance chez les philosophes et psychologues dès la fin du 19^e siècle, il tomba en désuétude au 20^e siècle avec l'avènement du courant béhavioriste (ou comportementaliste) qui ne s'intéressait qu'au comportement des individus et pas aux mécanismes internes de leur psychisme. Ce n'est qu'avec l'émergence et le succès des nouvelles approches cognitivistes dans la deuxième moitié du 20^e siècle que l'attention revint sur le devant de la scène.

1.1.1 Définitions

1.1.1.1 Les différents types d'attention

Le terme d'attention est utilisé en psychologie cognitive pour désigner une variété de sous-procédés qui ne concernent pas tous les travaux rapportés dans cette thèse. Le modèle de Sohlberg and Mateer (1987), issu d'études en neuropsychologie, vise à expliquer chacun de ces types d'attention qu'il recense :

— l'attention partagée
— l'attention alternée
— l'attention sélective
— l'attention soutenue
— l'éveil, ou arousal

Ainsi, au volant de sa voiture, un conducteur doit rester en éveil (arousal) et sélectionner les informations de la route sur lesquels se concentrer tout en ignorant les distractions (attention sélective). Il peut mener une conversation tout en prenant ses informations visuelles (attention partagée), doit savoir rester attentif sur une longue durée (attention soutenue) et peut déplacer son centre d'attention entre les voitures devant lui et le compteur de vitesse (attention alternée).

Chaque sous-procédé correspond à des mécanismes différents, relève de théories différentes et est étudié au moyen de méthodes spécifiques. Certains se recoupent parfois et il est peut donc être difficile d'établir une catégorisation tranchée des mécanismes mis en jeu par chacun. Nous ne pouvons néanmoins pas traiter d'un bloc les différents procédés regroupés sous le terme assez général d'attention. Le sujet qui nous intéresse dans cette thèse est celui d'*attention sélective*, autrement dit la capacité à sélectionner les informations que nous souhaitons traiter plus en profondeur et à ignorer les autres. En effet, nous verrons en partie 1.3 que la saillance est définie comme une composante particulière de cette forme d'attention. Nous utiliserons donc par la la suite le terme "attention" de manière récurrente pour désigner cette

forme d'attention, dite sélective.

1.1.1.2 L'attention sélective

James (1890) définit en 1890 l'attention comme suit [1] :
"*Everyone knows what attention is. It is the taking possession of the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalisation, concentration, of consciousness are of its essence. It implies withdrawal from some things in order to deal effectively with others, and is a condition which has a real opposite in the confused, dazed, scatterbrained state which in French is called distraction, and Zerstreutheit in German.*"

L'idée d'une forme de sélection est déjà présente dans cette première définition : c'est l'attention qui permettrait de sélectionner des stimuli sur lesquels on souhaite se concentrer et d'en ignorer d'autres. Un siècle plus tard, cette idée est toujours la même dans une revue sur l'attention visuelle réalisée par Yantis (2000) :
"People are perceptually selective : they subjectively experience and respond to only a subset of the sensory signals evoked by objects and events in the local environment. The psychological and neural mechanisms that mediate perceptual selectivity are collectively termed attention. Although often used to refer to other psychological phenomena (e.g., the ability to perform two or more tasks at the same time, or the ability to remain alert for long periods or time), for the purposes of this chapter, "attention" shall refer exclusively to perceptual selectivity."

Enfin, Kaya and Elhilali (2017), dans leur revue sur la modélisation de l'attention auditive, décrivent l'attention comme le médiateur entre la perception et le comportement, qui concentre les ressources sensorielles et cognitives sur les informations pertinentes dans l'espace des stimuli. Dans le cas de l'attention auditive, il s'agit donc d'un processus sélectif qui concentre ces

1. p 403-404

ressources sur les événements les plus pertinents d'une scène sonore.

Ces définitions, à des époques et dans des modalités différentes, reprennent l'idée de sélection : se concentrer sur un objet particulier d'une scène complexe implique d'en ignorer d'autres, et ce qu'on appelle l'attention est l'ensemble des mécanismes qui permettent de réaliser cette sélection. La focalisation de l'attention est donc le résultat d'une compétition entre différents objets. Mais elle peut également être modulée par les intentions du sujet, comme nous le présentons dans la partie suivante.

1.1.2 Attention descendante et ascendante

James (1890) distingue deux formes d'attention : l'attention active et l'attention passive. Cette distinction est aujourd'hui couramment admise sous une autre dénomination, l'attention ascendante (passive) et l'attention descendante (active) :

- lorsque l'individu dirige volontairement son attention sur une partie des stimuli qu'il reçoit en fonction de ses objectifs. C'est ce que fait par exemple un individu qui essaye de se concentrer sur une conversation spécifique dans un environnement bruyant. On parle alors d'attention descendante (ou "top-down").

- lorsque l'attention de l'individu est attirée par un stimulus indépendamment de sa volonté. Une alarme incendie qui se déclenche soudainement attire l'attention de l'auditeur, peu importe ses intentions et ce qu'il souhaite écouter. On parle alors d'attention ascendante (ou "bottom-up ").

Un exemple célèbre dans le domaine de l'audition permet de souligner la différence entre ces deux aspects : il s'agit de l'"effet cocktail party", introduit par Cherry (1953). Dans une réception, de nombreuses sources sonores assaillent le système auditif des individus : conversations, tintements de verres, musique de fond, etc. Cependant, ces derniers parviennent à analyser

leur environnement en orientant leur attention vers les informations qu'ils trouvent pertinentes, et ce malgré toutes les distractions. De nombreux effets sont certes mis en jeu (familiarité avec une voix écoutée, interaction avec la modalité visuelle). Mais de manière générale, ils peuvent se concentrer volontairement sur une conversation qui les intéressent (par des mécanismes descendants), ou être alertés par le bruit d'un verre qui se brise sur le sol (par des mécanismes ascendants).

1.1.3 Théories sur l'attention

Les propositions théoriques concernant la sélection attentionnelle ont fait l'objet d'une revue détaillée par Driver (2001). Nous en rappelons ici les principaux éléments.

Cherry (1953) étudia le premier le problème du "cocktail party" en se demandant comment les personnes étaient capables de suivre une conversation alors que plusieurs personnes parlaient en même temps. Il comprit que les caractéristiques physiques des voix perçues (sexe de l'interlocuteur, localisation, intensité de la voix) étaient utilisées par l'auditeur pour parvenir à ne prêter attention qu'à l'une d'entre elles. Dans son expérience, deux messages étaient diffusés simultanément dans chaque oreille des participants, avec la même voix et les mêmes caractéristiques physiques. Les participants n'étaient, dès lors, plus capables de séparer les deux messages. Par ailleurs, il montra que les informations sur lesquelles l'attention ne se concentre pas ne sont presque pas traitées. En effet, lorsque des participants devaient répéter un message parmi deux diffusés simultanément en écoute dichotique (un message différent dans chaque oreille), les participants ne remarquaient quasiment jamais quand le deuxième message, celui ignoré, était diffusé dans une autre langue ou à l'envers. Moray (1959) confirma ce résultat en montrant que des messages ignorés ne rentraient quasiment pas en mémoire même en ayant été diffusés 35 fois aux participants.

1.1.3.1 Le filtre attentionnel précoce

Sur la base des résultats de Cherry, Broadbent (1958) proposa une théorie de l'attention autour de l'idée d'un filtre précoce. Ce dernier opèrerait sur la multitude de stimuli physiques qui parviennent à nos organes sensoriels simultanément et qui seraient conservés dans une mémoire immédiate, dite sensorielle, avec leurs caractéristiques physiques. Le filtre réaliserait alors la sélection sur la base des propriétés physiques mémorisées de ces informations. On parle de théorie du filtre précoce de l'attention. Cette théorie explique bien les résultats trouvés par Cherry (1953). Cependant, elle suppose que les informations auxquelles on ne prête pas attention ne sont jamais traités plus en profondeur, ce qui fut par la suite remis en question. Underwood (1974) montra notamment que des informations ignorées pouvaient être détectées malgré tout, et ce d'autant plus si les participants étaient expérimentés. Allport et al. (1972) révélèrent l'importance de la similarité entre les informations à sélectionner : si les messages à ignorer ressemblaient au message à retenir, comme dans la tâche de Cherry (1953), il s'avérait plus difficile de les traiter également. En revanche, les participants retenaient très bien en mémoire des informations à ignorer si elles étaient suffisamment différentes de l'information à laquelle ils devaient faire attention : des images avec des mots par exemple.

Ces résultats ne peuvent être expliqués par la théorie du filtre attentionnel précoce de Broadbent. De fait, son système de sélection trop rigide ne permet pas d'expliquer la variabilité observée dans le degré de traitement des informations à ignorer.

1.1.3.2 Le filtre attentionnel atténué

Treisman (1960) reproduisit la tâche réalisée par Broadbent (1958) et observa que dans une faible proportion des cas (6%), les participants rappelaient le mot qu'ils devaient ignorer. Ce phénomène avait lieu lorsque le mot à ignorer était plus probable, dans le contexte, que celui à retenir. Treisman (1964) proposa donc une théorie, adaptée de celle de Broadbent, dans laquelle le

filtre attentionnel atténue le signal des informations à ignorer. Dans cette théorie, les informations ont un seuil d'activation dont le niveau dépend de la quantité d'attention nécessaire pour les identifier : plus une information est attendue dans un contexte donnée, plus son niveau d'activation est bas, et inversement. Des informations non attendues deviennent alors potentiellement identifiables si leur seuil d'activation est suffisamment bas.

Treisman parvint donc à expliquer les résultats dont la théorie de Broadbent ne pouvait pas rendre compte. De fait, avec sa proposition de filtre atténué, une partie des stimuli non attendus peut finalement être traitée en profondeur.

1.1.3.3 Le filtre attentionnel tardif

Deutsch and Deutsch (1963) approfondirent encore cette idée en suggérant que la totalité des stimuli est traitée en profondeur. Ils furent les premiers à proposer une théorie du filtre attentionnel tardif. Selon eux, toutes les informations bénéficieraient d'un traitement approfondi, et la sélection opérerait seulement après le stade de l'analyse sémantique. Cette théorie explique également les résultats non expliqués par le filtre précoce de Broadbent.

Les théories sur le filtre attentionnel sont résumées en figure 1.1.

1.1.3.4 Évaluation de ces propositions théoriques et théorie actuelle

Si la théorie du filtre attentionnel précoce fut rapidement limitée, celles de Treisman (1960) et de Deutsch and Deutsch (1963) proposaient toutes les deux une explication raisonnable aux résultats dont Broadbent (1958) ne rendait pas bien compte. Une série d'études utilisant le même paradigme d'écoute dichotique de messages vocaux dont l'un des deux devait être ignoré et dans lesquels les participants devaient détecter des cibles (Deutsch et al., 1967 ; Treisman and Geffen, 1967 ; Treisman and Riley, 1969) opposa les arguments des deux théories candidates mais ne permit pas de trancher sur le positionnement du filtre attentionnel.

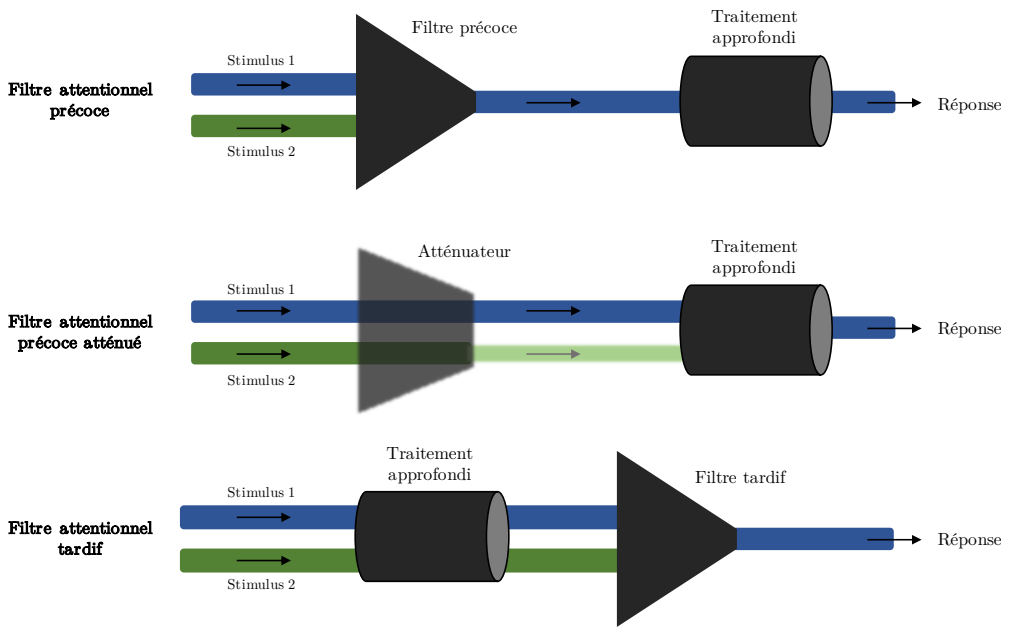

FIGURE 1.1 – Les théories de filtres attentionnels. Le filtre attentionnel précoce de Broadbent, la version atténuée de Treisman et le filtre tardif de Deutsch et Deutsch.

Il fallut attendre l'arrivée d'études en neurosciences pour pouvoir confronter ces deux théories. Woldorff et al. (1993) enregistrèrent les potentiels évoqués (ou "Event-Related Potentials", ERP) dans le cortex auditif par des stimuli attendus et des stimuli ignorés. Ils révélèrent qu'entre 20 et 50 ms après la présentation des stimuli les ERP étaient plus intenses pour les stimuli attendus.

Autrement dit, l'activation initiale du cortex auditif est plus grande pour les stimuli sur lesquels l'attention est focalisée que pour ceux qui sont ignorés, ce qui témoigne d'un filtrage précoce de l'information par les processus attentionnels.

Lachter et al. (2004) confirmèrent ce résultat en montrant que des stimuli situés hors de la zone d'attention ne bénéficiaient pas d'identification lexicale. Dans leur expérience, des amorces sémantiques pouvaient apparaître dans une zone identique à celle de la cible sur laquelle les participants devaient prendre une décision lexicale, ou en dehors de cette zone. Les effets

d'amorce se révélèrent nuls lorsque le stimulus n'était pas dans la zone de focus attentionnel des participants.

Ainsi, les stimuli sur lesquels l'attention n'est pas focalisée ne semblent pas bénéficier d'un traitement approfondi contrairement à ce que la théorie du filtre attentionnel tardif prévoyait.

Les résultats actuels convergent donc vers une théorie du filtre attentionnel précoce, que la proposition de Treisman (1960) assoit comme la plus à même d'expliquer les résultats observés.

Treisman and Gelade (1980) pousseront cette proposition plus loin avec la théorie de l'intégration des caractéristiques qui tente de rendre compte de ce qui précède le filtre attentionnel. Selon celle-ci, en situation d'exposition à de multiples stimuli, des informations sur différentes caractéristiques sont recueillies à un stade pré-attentif. Les caractéristiques sont stockées dans une carte : chaque emplacement de la carte, où des caractéristiques ont été détectées, donne accès aux caractéristiques correspondantes. Lorsque l'attention est portée sur un emplacement donné de la carte, les caractéristiques sont alors prises en compte et stockées comme un objet, qui peut ensuite être identifié. Cette deuxième étape est appelée "stade de l'attention focalisée". Cette théorie a certes été développée spécifiquement pour la modalité visuelle, mais elle a été le concept à partir duquel les premiers modèles de saillance ont été développés dans la modalité auditive (voir 1.2.2).

Ainsi, il est communément admis à ce jour que les propriétés physiques des stimuli qui nous parviennent sont traitées à un niveau pré-attentionnel, puis que le filtre attentionnel intervient pour atténuer les signaux à ignorer (voir partie 1.1.3). Il y a bien atténuation et non pas extinction de ces signaux : il pourrait arriver que des stimuli non attendus parviennent à franchir ce filtre attentionnel, indépendamment des intentions de l'auditeur. On parle alors de capture attentionnelle. C'est cette capacité que nous appelons saillance et à laquelle nous nous intéressons maintenant.

EN RÉSUMÉ

— Si le terme d'attention est largement utilisé et peut désigner des procédés de natures différentes en psychologie cognitive, nous la définissons dans ces travaux comme la capacité à sélectionner les informations pertinentes et à ignorer les autres. Autrement dit, nous appelons ici attention ce qui est parfois plus spécifiquement désigné par l'appellation *attention sélective*.

— Différentes théories ont été proposées pour expliquer comment les stimuli sélectionnés ou ignorés sont filtrés après avoir été perçus. À ce jour, la théorie du filtre attentionnel précoce et atténuateur est la plus à même d'expliquer les résultats expérimentaux.

— Dans le cadre de cette théorie, les propriétés physiques de stimuli non attendus peuvent franchir le filtre attentionnel : il est ainsi possible que des stimuli non-pertinents capturent l'attention d'un individu contre sa volonté. On parle alors de capture attentionnelle.

1.2 La saillance

1.2.1 Définition

L'utilisation du terme "saillant" peut être trompeuse, car il est souvent employé pour parler d'un stimulus lui-même. Or, nos organes sensoriels ne reçoivent jamais des informations isolées. Pourrait-on alors parler de la saillance d'un son en soi ? On se convainc par exemple aisément qu'un cri de bébé qualifié de saillant dans une ambiance sonore de bureau le serait beaucoup moins dans celle d'une crèche. Par ailleurs, dans le silence, n'importe quel son serait saillant. On ne peut donc pas parler de saillance sans prendre en compte le contexte ou le fond sonore.
C'est d'ailleurs l'une des idées principales que l'on retrouve derrière le terme tel qu'il est utilisé dans la littérature. Un son est dit saillant quand il *ressort*, se dégage, se distingue aisément, dans une scène sonore (De Coensel and Botteldooren, 2010 ; Liao et al., 2016).
D'autres auteurs ont utilisé ce terme pour évoquer plus spécifiquement le mécanisme qui permet au stimulus de se distinguer du reste : un son saillant est un son qui contraste avec le fond sonore, qui crée une *divergence* dans l'information auditive et les régularités qu'elle contient (Tordini et al., 2016 ; Tsuchida and Cottrell, 2012).
D'autres enfin utilisent ce terme pour évoquer la finalité, la conséquence au niveau cognitif : un son saillant est un son qui capture l'attention d'un auditeur (Itti and Koch, 2001 ; Kaya et al., 2020 ; Tsiami et al., 2016 ; Zhao et al., 2019). Autrement dit, la saillance d'un son est sa capacité à attirer l'attention en mettant en jeu des mécanismes bottom-up.

Ces différents usages tournent tous autour de la même idée. Un stimulus *capture l'attention* parce qu'il *ressort* dans le fond sonore, et cela arrive sans doute parce qu'il crée une *divergence* dans la régularité des informations perçues. Le terme "saillant" est alors simplement utilisé pour décrire le même phénomène, mais à différents niveaux d'interprétation.
Or, nous ne savons pas exactement pourquoi et comment notre système cog-

nitif sélectionne un stimulus plutôt qu'un autre. Ainsi, la définition la plus "prudente", c'est-à-dire qui ne fait pas d'hypothèse sur un mécanisme ou un autre, est bien celle qui part de la finalité du processus. Un stimulus est dit saillant lorsque il s'impose à notre attention. Les questionnements autour du "pourquoi" et du "comment" ne rentrent pas dans notre définition ici, que nous gardons comme la plus conservatrice.
Retenons ainsi la définition suivante :

*Un stimulus est dit **saillant** lorsqu'il est susceptible de capturer l'attention d'un individu indépendamment de sa volonté (de manière "bottom-up") dans un certain contexte.*

C'est ce choix d'assimiler la notion de saillance à celle de composante bottom-up de l'attention qui a d'ailleurs majoritairement été fait dans la littérature, notamment dans les travaux de modélisation de ce processus (voir Kaya and Elhilali (2017) pour une revue de littérature).

1.2.2 Déterminants de la saillance auditive

L'un des axes de recherche sur la saillance consiste à essayer de comprendre quelles caractéristiques des stimuli sont susceptibles de les rendre saillants. Nous parlons ici de propriétés sonores ou perceptives bas-niveaux. On sait par exemple que la formulation de son prénom (Wood and Cowan, 1995) ou que des sons liés à de fortes émotions (Vuilleumier, 2005) sont susceptibles d'être saillants, mais ces résultats sortent de notre périmètre d'intérêt.

Dans la modalité auditive, la sonie est le premier paramètre psychoacoustique qui vient à l'esprit quand on parle de saillance : plus un son est perçu comme fort, plus il a de chance de capturer l'attention. Son importance n'est ainsi plus à démontrer (Huang et al., 2017 ; Liao et al., 2016 ; Tordini et al., 2016). La question qui se pose alors est la suivante : si des sons sont de même

sonie, comment l'un d'eux peut être rendu plus saillant que les autres ?

La réponse à cette question n'est pas encore définitivement formulée, mais des pistes de recherche existent. Un effet de la hauteur (Dalton and Lavie, 2004 ; Kaya and Elhilali, 2014) a par exemple déjà été relevé. Dalton and Lavie (2004) ont ainsi mis en évidence, dans leur paradigme (voir partie 2.1.1.2), que la présence d'un son de hauteur différente dans une succession de sons ayant tous une hauteur identique induisait une augmentation des temps de réponse des participants. Kaya et al. (2020) ont également montré que des variations de hauteur de certaines notes au sein d'une mélodie pouvaient moduler la réponse neuronale induite par ces notes. Leur étude montre que le timbre peut également jouer un rôle dans la modulation de la saillance. Cette idée confirme les résultats de Tordini et al. (2016) qui ont montré que, pour des gazouillements d'oiseaux, la brillance avait une contribution significative dans la prédiction de leur saillance. Par ailleurs, le rôle de la rugosité a été souligné dans une étude de Zhao et al. (2019), qui ont trouvé une corrélation significative entre rugosité et saillance, confirmant les travaux d'Arnal et al. (2015).

La quête visant à mettre en évidence les déterminants acoustiques de la saillance auditive est motivée par une finalité : la prédiction de la saillance dans une scène sonore. Ainsi, on retrouve les éléments mentionnés ici (sonie, brillance, rugosité) dans la conception des modèles de saillance.

1.2.3 Modéliser la saillance auditive

La question de la saillance a d'abord été principalement étudiée dans la modalité visuelle. Il n'est donc pas surprenant que les premiers modèles de saillance auditive se soient inspirés des modèles de saillance visuelle.

Les premiers, et notamment celui de Kayser et al. (2005) étaient basés sur la théorie de l'intégration des caractéristiques décrite plus haut (voir 1.1.3.4). L'idée initiale était de considérer une scène sonore comme une image en

deux dimensions : temps et fréquence (on parle de "spectrogramme"). Le modèle extrayait ensuite certaines caractéristiques de cette image : contrastes d'intensité, de temps et de fréquence. Les caractéristiques étaient extraites à différentes échelles, et permettaient de générer des cartes, qui étaient ensuite normalisées et intégrées pour obtenir une carte de saillance auditive. Le résultat était donc une carte dans l'espace temps-fréquence, donnant une valeur de saillance au signal à tout instant et pour chaque fréquence.

Puis, d'autres travaux basés sur cette architecture furent menés pour prendre en compte davantage de caractéristiques et améliorer leur analyse. Kalinli and Narayanan (2007) ajoutèrent deux caractéristiques d'orientation et une caractéristique de hauteur. Duangudom and Anderson (2007) considérèrent des caractéristiques plus spécifiques des processus de la perception auditive, en particulier les modulations spectrales et temporelles. De plus, l'analyse des caractéristiques fut améliorée et permit la dérivation de ces caractéristiques pour obtenir une cartographie du signal à plusieurs échelles, plus précise que les cartes des modèles précédents.

Cependant, ces modèles souffraient du même défaut : en considérant le signal audio comme une image bidimensionnelle, on perdait l'information qui provient du fait que le signal est perçu dans un certain ordre temporel. Sur une image, il n'y a pas de contrainte d'ordre pour parcourir l'image ; mais un signal audio est reçu dans le temps, ce qui correspond à un parcours de gauche à droite sur le spectrogramme. Ainsi, le résultat de la carte de saillance devait être différent si l'on parcourait l'image dans les deux sens, ce qui n'était pas le cas avec les premiers modèles issus de la modalité visuelle.
En d'autres termes, la représentation bidimensionnelle du spectrogramme ne peut être simplement traitée comme une image car l'une des deux dimensions, le temps, joue un rôle particulier dans la perception auditive.

Ainsi, la modélisation de l'attention auditive nécessite de considérer les caractéristiques comme des éléments variant dans le temps et dont les variations jouent un rôle clé. La première étude à considérer cette approche fut

celle de Kaya and Elhilali (2012). Les caractéristiques furent cartographiées dans le temps, et non sur des images bidimensionnelles. Ces caractéristiques étaient d'ailleurs plus proches des caractéristiques perceptives du son que nous avons l'habitude de manipuler en psychoacoustique : hauteur, intensité sonore, timbre. Le résultat était donc une courbe de saillance temporelle.

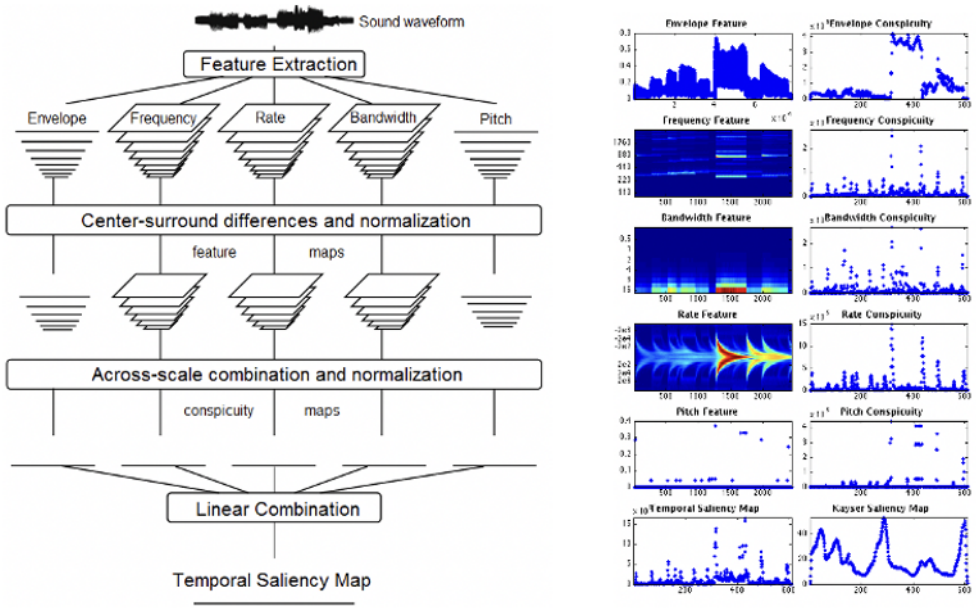

FIGURE 1.2 – Courbe de saillance auditive obtenue par Kaya and Elhilali (2012)

Kaya and Elhilali (2014) proposèrent une nouvelle piste, basée sur la théorie du codage prédictif et de la détection des divergences. Dans leur modèle, des caractéristiques comme l'intensité sonore, le timbre, la hauteur étaient collectées au fil du temps, et à chaque instance temporelle, une prédiction de la valeur suivante pour chaque caractéristique était faite. Les moments où les caractéristiques entrantes étaient trop éloignées des prédictions étaient signalés comme des moments saillants. De plus, une interaction non linéaire entre les caractéristiques fut ajoutée au modèle.

Kim et al. (2014) adoptèrent une approche différente, en utilisant des évaluations comportementales de saillance pour entraîner un classificateur linéaire optimisant la séparation entre les événements saillants et non saillants. Il apparut que le classificateur utilisait les contrastes temporels et spectraux

pour maximiser ses performances. Au même moment, Tordini et al. (2013) testaient la contribution de diverses caractéristiques à la saillance ; le centroïde temporel, le centroïde spectral, l'harmonicité, la durée effective et le tempo s'avéraient corrélés aux évaluations de saillance.

Schauerte and Stiefelhagen (2013) abordèrent ce problème de manière légèrement différente, en utilisant la notion de "surprise", évaluée par des méthodes statistiques (plus la probabilité qu'un événement sonore se produise à un moment donné est faible, plus la surprise est grande). Ce modèle fut ensuite amélioré par Rodriguez-Hidalgo et al. (2018) chez lesquels le cochléogramme était considéré pour calculer la surprise pour chaque bande de fréquence à différentes échelles de temps. Les différentes échelles étaient ensuite recombinées pour obtenir la surprise moyenne en fonction du temps.

Enfin, et plus récemment, Huang et al. (2017) puis Kothinti et al. (2021) ont testé la contribution d'un large ensemble de caractéristiques à la saillance sur un ensemble de scènes sonores environnementales variées. Des caractéristiques comme la sonie, l'harmonicité, la hauteur, la brillance ou les modulations temporelles se sont révélées avoir des rôles déterminants dans le modèle pour expliquer les évaluations subjectives de saillance. Ce modèle étant utilisé dans la suite de nos travaux, il sera plus détaillé au chapitre 4.

Dans l'ensemble, si ces travaux semblent de plus en plus s'orienter vers une modélisation de la saillance auditive spécifique à sa modalité perceptive, les mécanismes sous-jacents et fondamentaux propres à la question de l'attention auditive ne sont pas encore tous compris et donc intégrés dans les modèles.

Or, pour pouvoir améliorer les modèles existants, il faut d'une part un algorithme pertinent fondé sur des mécanismes propres au traitement cognitif de l'audition, et d'autre part des données permettant d'entraîner cet algorithme. L'étude de la saillance auditive repose ainsi sur la capacité à en réaliser une mesure expérimentale.

1.2.4 Mesurer la saillance

Déterminer la saillance d'un stimulus dans un environnement donné, découvrir et quantifier l'impact de nouveaux déterminants de la saillance ou modéliser et prédire la saillance dans une scène sonore nécessitent une ressource commune : des données de saillance réelles, mesurées sur des individus. Autrement dit, on ne peut étudier la saillance auditive en s'affranchissant de la collecte de mesures subjectives de saillance.

Ce sujet fait l'objet du chapitre 2, dans lequel nous mettrons en oeuvre un paradigme de mesure de la capture attentionnelle pour révéler la composante bottom-up de l'attention et la manière dont elle est modulée par les attributs du timbre. Pour choisir ce paradigme, nous avons du mener une revue exhaustive des différentes méthodes expérimentales envisageables. Un état de l'art sur ce sujet est donc disponible en partie 2.1. Nous y distinguons notamment deux méthodes distinctes de mise en évidence d'un phénomène de capture attentionnelle : la mesure explicite et la mesure implicite.

Dans une mesure de capture attentionnelle explicite, on demande aux participants de réaliser une tâche et on observe si la présence d'un stimulus non attendu est remarquée. L'exemple le plus connu est celui de Simons and Chabris (1999) [2]. Deux équipes de joueurs de basket-ball s'échangeant une balle sont présentées, l'une en blanc, l'autre en noir. Les participants doivent compter le nombre de passes réalisées entre les joueurs de l'équipe blanche. Au cours de la scène, une personne déguisée en gorille, noir, traverse la scène (voir figure 1.3). Lorsqu'on demande à la fin de la tâche aux participants s'ils ont remarqué quelque chose de particulier, plus de la moitié n'a pas remarqué le gorille. On parle de *cécité attentionnelle*. Pour les autres participants (ceux ayant remarqué le gorille alors que leur attention était focalisée sur les passes de l'équipe blanche), il y a eu capture attentionnelle. La mesure de capture attentionnelle peut donc s'estimer avec la proportion de participants dans ce dernier cas.

2. https://www.youtube.com/watch?v=vJG698U2Mvo&themeRefresh=1

FIGURE 1.3 – Expérience de Simons and Chabris (1999). Plus de la moitié des participants occupés à compter les passes des joueurs blancs ne remarquent pas le gorille.

Dans une mesure de capture attentionnelle implicite, on observe la dégradation de la performance des participants dans une tâche donnée en fonction de la présence de stimuli non-pertinents pour la réalisation de cette tâche. Le paradigme que nous détaillerons au chapitre 2, intitulé paradigme du singleton additionnel, en est un bon exemple. Un exemple générique est proposé en figure 1.4 : la tâche est de retrouver le rond parmi les carrés. Le stimulus non pertinent, un carré d'une autre couleur appelé *singleton*, peut parfois être présent. Si on mesure le temps de réaction des participants pour détecter le rond, on remarque qu'il est en moyenne plus important lorsque le singleton est présent : on met ainsi en évidence la capture attentionnelle par le singleton.

Les méthodes implicites se prêtent plus à notre étude, car elles permettent de quantifier la capture attentionnelle de manière plus précise (dégradation de performance pour chaque participant, par rapport à une simple proportion de cécité attentionnelle dans les méthodes explicites) et sont plus facilement compatibles avec des modulations de caractéristiques des stimuli dont on souhaite mesurer la saillance. Ces aspects sont plus détaillés en partie 2.1, où nous revenons sur les méthodes envisageables et motivons le choix que avons fait pour notre étude.

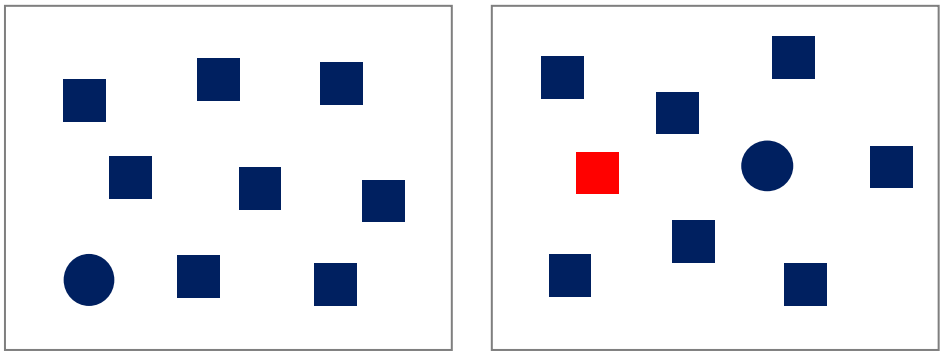

FIGURE 1.4 – Exemple générique de recherche d'une cible (le rond) parmi des distracteurs (les carrés) dans deux cas : à gauche, en absence de singleton, à droite, en présence d'un singleton (le carré rouge).

1.2.5 Mise en évidence de la composante bottom-up de l'attention

Le principe des modèles de saillance tel que rapporté en section 1.2.3 repose sur l'encodage de propriétés sonores et la combinaison de leurs variations temporelles pour prédire la saillance d'une scène sonore à chaque instant. Ce principe suppose donc que la composante bottom-up de l'attention est modulée par les variations des caractéristiques acoustiques et/ou psychoacoustiques de l'environnement sonore perçu. On dit alors que la capture attentionnelle est *stimulus-driven*. Or, la modulation d'un effet de capture attentionnelle par les propriétés physiques des stimuli n'a pas encore été mise en évidence expérimentalement. Ce sera l'objet du chapitre 2.

Ce sujet a longtemps fait débat dans la modalité visuelle, où la capture attentionnelle a largement été étudiée (voir Simons (2000) pour une revue). Les approches implicites entre autres (voir partie 1.2.4) ont permis de montrer que des stimuli non pertinents définis par leur couleur, leur forme ou leur moment d'apparition captent l'attention de participants effectuant une tâche de recherche visuelle (Theeuwes, 1992,9 ; Yantis and Jonides, 1984).

Toutefois, une question est restée ouverte : est-ce que les stimuli non-pertinents capturent l'attention à cause de leurs caractéristiques, ou est-ce

que les participants sont dans un état attentionnel qui les rend susceptible de porter leur attention sur ces stimuli ? Certains ont soutenu que les objets saillants captent automatiquement l'attention, quels que soient les objectifs induits par la tâche menée par l'individu. Ils ont observé que certaines caractéristiques, telles que la couleur ou la forme, permettent à l'objet saillant d'attirer automatiquement l'attention (Theeuwes, 2010). Cela a conduit à une conception de la capture attentionnelle dite *stimulus-driven* (Theeuwes, 1993) : la sélection visuelle est déterminée par les caractéristiques physiques des stimuli et l'attention est attirée vers l'endroit où un objet est différent des autres sur une dimension particulière. D'autres, en revanche, ont soutenu que seuls les éléments ayant des caractéristiques les rendant similaires à la cible recherchée peuvent capter l'attention. Pour eux, la capture ne dépendrait que de l'état attentionnel de l'individu conditionné par la tâche, et donc des mécanismes top-down qu'il déploierait (Folk et al., 1992) : elle est alors dite *contingente*.

Les auteurs des différentes parties se sont finalement réunis pour examiner leurs propositions ensemble et ont pu accorder leurs théories (Luck et al., 2021). Ils ont convenu que "les stimuli physiquement saillants génèrent automatiquement un signal de priorité qui, en l'absence de paramètres de contrôle attentionnel spécifiques, capte automatiquement l'attention, mais qu'il existe des circonstances dans lesquelles la capture effective de l'attention peut être évitée", réconciliant ainsi les approches de la capture guidée par le stimulus et de la capture contingente.

Ainsi, l'existence d'une composante bottom-up de l'attention, dite *stimulus-driven*, est aujourd'hui communément admise. Mais à notre connaissance, il n'y a jamais eu de preuve explicite d'une relation entre les variations d'une caractéristique du stimulus et l'effet de capture attentionnelle qu'il génère. En d'autres termes, le pilotage de l'effet *stimulus-driven* n'a jamais été mis en évidence et quantifié : nous ne savons pas comment l'effet de capture attentionnelle évolue avec les variations d'une caractéristique donnée.

Les paradigmes évoqués en 1.2.4 et ceux ayant nourri les débats rapportés en 1.2.5, propres à la psychophysique, mettent en oeuvre des mesures de capture attentionnelle dans des scènes très simples. Quelques figures géométriques ou quelques sons simples sont présentés aux participants, dont la tâche consiste à retrouver une cible. Or, les motivations de ce travail ont puisé leur source dans l'étude de l'exposition à des scènes sonores environnementales. Ainsi, l'une des visées de ces travaux est de monter en complexité en termes de scènes sonores et de mécanismes cognitifs en jeu et d'observer les effets de la saillance dans cette complexité croissante. Cependant, la perception de scènes sonores complexes met en jeu des mécanismes variés et des principes d'analyse qu'il convient de préciser à ce stade.

EN RÉSUMÉ

— La saillance est ici définie comme la composante bottom-up de l'attention, c'est-à-dire la capacité d'un stimulus à capturer l'attention d'un individu indépendamment de sa volonté.

— Certaines caractéristiques acoustiques/psychoacoustiques sont susceptibles de rendre un son saillant. Des pistes de déterminants de la saillance auditive émergent : sonie, hauteur ou attributs du timbre (brillance, rugosité).

— Les modèles de prédiction de saillance auditive encodent les variations temporelles de ces caractéristiques acoustiques et psychoacoustiques et en déduisent une prédiction de saillance en fonction du temps.

— Différentes méthodes expérimentales sont envisageables pour mesurer un effet de capture attentionnelle. Nous choisirons au chapitre 2 un paradigme en particulier et nous détaillerons les motivations de ce choix.

— Il n'y a à ce jour pas de preuve expérimentale qui ait mis en évidence la relation entre les variations des propriétés d'un son et sa capacité à capturer l'attention.

1.3 La perception de scènes complexes

1.3.1 Analyse de scènes complexes

Une immense quantité d'informations réparties dans le temps et l'espace, variables voire parfois partiellement occultées, assaillent nos sens en permanence. Notre cerveau organise à chaque instant cette foule de sollicitations sensorielles et la convertit en une scène cohérente faite d'objets sensés (Bizley and Cohen, 2013). En vision, Marr (1982) a conceptualisé les principes régissant l'analyse de scènes complexes. Dans la modalité auditive, Bregman (1994) a caractérisé l'ensemble des principes d'analyse des stimulations sonores sous l'appellation d'Analyse de Scènes Auditives (ASA). Un de ces principes consiste par exemple à déterminer si un ensemble d'informations sonores provient d'une seule ou de plusieurs sources, et s'il convient donc de le traiter comme un seul ou plusieurs flux (Itatani and Klump, 2017). La compréhension des mécanismes en jeu et de la manière dont l'analyse des scènes visuelles ou auditives complexes est réalisée reste cependant un défi majeur en sciences cognitives aujourd'hui (Cichy and Teng, 2017 ; Kaya and Elhilali, 2017). Kondo et al. (2017) ont proposé une revue de l'état de la recherche sur l'analyse de scènes auditives et visuelles.

1.3.1.1 Influence de mécanismes ascendants et descendants dans l'Analyse de Scènes Auditives

Les principes de l'Analyse de Scène Auditive (ASA) mettent inévitablement en jeu des caractéristiques du signal acoustique, et donc les mécanismes perceptifs ascendants associés. Un point central de l'ASA consiste par exemple à traiter l'information issue d'une mixture auditive pour l'organiser en une scène cohérente faite de différents objets, qui peuvent ensuite être perçus en tant que tel. L'étude de la ségrégation séquentielle de flux auditifs entremêlés a été largement menée depuis l'introduction les travaux de Noorden (1975). On sait aujourd'hui que de nombreuses caractéristiques acoustiques sont susceptibles de favoriser la ségrégation des flux (Grimault, 2013 ; Moore

and Gockel, 2002) : hauteur, sonie, bande passante, fréquences notamment. La ségrégation de sources simultanées, notamment étudiée dans le cadre de tests d'identification de voyelles présentées simultanément, se trouve également influencée par les propriétés acoustiques des signaux, entre autres leur différence de hauteur (de Cheveigné, 1999).

Cependant, la perception d'une scène complexe met simultanément en jeu des mécanismes à la fois ascendants et descendants (Kondo and Kashino, 2009 ; Veale et al., 2017). L'information ne remonte pas simplement depuis les stimuli vers l'individu en suivant des étapes successives pour construire étape par étape des représentations internes plus haut-niveaux. De fait, l'analyse d'une scène se fait toujours dans un certain contexte. Ce dernier peut être associé à des connaissances chez l'individu et donc influencer sa perception de la scène. Bregman (1994) a ainsi évoqué l'idée de schémas mémorisés qui permettraient de favoriser le repérage de cibles dans une mixture auditive. Depuis, cette proposition a été validée par l'expérience (Dowling et al., 1987) : des participants savent extraire des mélodies familières intercalées dans des mélodies non-familières même sans différence d'indices connus pour favoriser la ségrégation (voir paragraphe précédent). De plus, Devergie et al. (2010) ont montré que leurs performances pour extraire la mélodie familière étaient modulées par un mécanisme attentionnel focalisé sur le rythme des mélodies. Cela prouve que les connaissances préalables peuvent influencer l'activation de mécanismes descendants permettant de favoriser l'analyse d'une scène auditive.

L'influence de divers procédés top-down a aujourd'hui été mise en évidence : attention descendante, intentions de l'individu, connaissances préalables (Kaya and Elhilali, 2017 ; Moore, 2012 ; Snyder et al., 2012). Les interactions pouvant exister entre ces mécanismes et les mécanismes ascendants, ainsi que leur influence relative, sont néanmoins encore incomprises (Kondo et al., 2017).

Ainsi, la mise en oeuvre de mécanismes ascendants et descendants permet d'organiser une scène en un ensemble cohérent constitué des différents objets auditifs que perçoit l'auditeur. Confronté à la perception de différentes sources au sein de la même scène, celui-ci doit mener le traitement de l'information sonore en situation de multi-exposition.

1.3.1.2 Perception sonore en situation de multi-exposition

En situation d'exposition à une scène complexe, l'individu reçoit des informations issues de différentes sources variées, simultanées ou non. Dans ce contexte, les sources concurrentes ne sont pas perçues comme une somme de sources isolées dont la perception globale résulterait de la perception de chacune si elle était isolée. Au contraire, les percepts interagissent de manière complexe. L'exemple de l'évaluation de la sonie permet de s'en rendre compte.

De fait, Susini et al. (2002) ont montré que la sonie globale d'une séquence sonore de quelques dizaines de secondes dépendait des évènements ayant un niveau sonore prédominant et de leur position dans la séquence. Ponsot et al. (2013) ont ensuite confirmé cet effet de dominance de niveau avec des sons variables en intensité. Plus récemment Vannier et al. (2018) ont observé que lorsque deux sources de sonie différentes étaient perçues simultanément, le jugement de sonie globale de la scène était dominé par la sonie de la source la plus saillante uniquement. Ces résultats sont dans la lignée de ceux de Kuwano and Namba (1985), qui ont proposé une mesure de prédiction de l'évaluation de l'impression globale du niveau dans des séquences urbaines en ne prenant en compte que les évènements acoustiques au niveau élevé. Fastl (1991) a confirmé cette prédominance des évènements les plus bruyants en montrant que le jugement de sonie globale était le mieux prédit par une valeur de sonie atteinte pendant seulement 4% du temps.

La perception d'une scène contenant de multiples sources, simultanées ou non, implique ainsi des interactions complexes entre les percepts liés à chaque source. Les évaluations de jugements de sonie ici montrent en effet

que dans certaines situations multi-sources, seules quelques unes d'entre elles ont une influence sur la perception globale de la scène. Ces résultats ne sont pas propres à la sonie : Powell (1979) a ainsi montré que la gêne causée par une source sonore pouvait être inhibée en présence d'une autre source plus gênante. Alayrac et al. (2011) ou Morel et al. (2012) ont ensuite confirmé que la gêne causée par une source seule n'était pas la même que la gêne causée par la même source dans une situation de multi-exposition.

L'importance du contexte de multi-exposition est ainsi primordiale : une source ne saurait être perçue indépendamment de ses sources concurrentes et donc de la perception globale de la scène. Cette perception globale d'une scène relativement à celle des multiples éléments qui la composent a été étudiée comme un objet à part entière dans le champ de la psychologie cognitive, et a permis de révéler un effet de primauté du traitement holistique.

1.3.2 Primauté du traitement holistique de l'information

Une composante clé de la perception, nécessairement impliquée dans la perception de scènes complexes, concerne les relations entre la perception d'un ensemble et celle de ses parties. Historiquement, deux écoles de pensée se sont affrontées sur ce sujet. D'une part, les structuralistes défendaient que tout ensemble sensoriel, c'est-à-dire toute perception d'un objet dans sa globalité, était le résultat de l'assemblage de toutes les sensations élémentaires liées à chacun des éléments qui le composent (Titchener, 1909 ; Wundt, 1893). D'autre part, les gestaltistes soutenaient que tout ensemble sensoriel était différent du simple assemblage de ses composantes élémentaires (Koffka, 1935 ; Köhler, 1967 ; Wertheimer, 1938). Pour eux, la perception de chaque élément était même, à l'inverse, influencée par l'ensemble dont il faisait partie. Le terme de traitement holistique est donc utilisé, dans l'esprit de celui de la théorie de la Gestalt, pour parler de la primauté du traitement des propriétés globales dans la perception. On utilisera ainsi de manière équivalente les termes de traitement holistique ou de traitement global.

Dans ce cadre, Navon (1977) a émis l'hypothèse de précédence globale, qui a rapidement été considérée comme une version moderne des théories des gestaltistes selon lesquelles le traitement global est réalisé en priorité lors de la perception d'objets visuels (Pomerantz, 2017 ; Robertson, 1986 ; Treisman, 1986 ; Uttal, 1988). Cette hypothèse a été le fondement d'études expérimentales focalisées sur la question de la hiérarchie du traitement local *vs.* global de l'information et que nous présentons maintenant.

1.3.2.1 Mise en évidence expérimentale de la hiérarchie du traitement local/global de l'information en vision

La primauté du traitement holistique, abordée en utilisant la métaphore de "forêt avant les arbres", est étudiée depuis longtemps dans le domaine visuel. Elle caractérise la façon dont nous traitons de manière différenciée les échelles locales et globales de l'information spatiale, la priorité étant donnée au traitement global de manière générale. Diverses expériences ont été menées en laboratoire pour mettre ce phénomène en évidence (voir la revue de Kimchi (1992)). On parle de *paradigme local/global* de manière générale pour évoquer ces méthodes expérimentales.

L'une des approches couramment adoptées et que nous décrirons plus en détail au chapitre 3 consiste à présenter des grands caractères composés de caractères plus petits et à analyser comment les petits caractères à l'échelle locale influencent la perception du caractère plus grand à l'échelle globale, et vice versa (voir figure 1.5). Navon (1977) a ainsi observé que les participants identifiaient plus rapidement et plus précisément les informations à l'échelle globale. De plus, lorsque les informations aux deux échelles étaient incongruentes, l'information globale perturbait davantage l'identification de l'information locale que l'inverse. Il a donc été conclu qu'il existait bien un traitement privilégié des informations à l'échelle globale, appelé "effet de précédence globale".

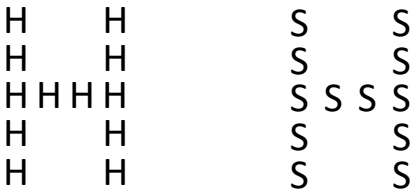

FIGURE 1.5 – Exemple de lettres utilisées dans le paradigme local/global. À gauche, les informations au échelles locales et globales sont congruentes, à droite, incongruentes.

Bien que cet effet soit constaté dans différentes conditions de présentation des stimuli et de tâche, il semble qu'il puisse être inversé dans certains cas. Mevorach et al. (2006) ont par exemple montré que le traitement local/global des informations spatiales était réorganisé par la saillance visuelle des stimuli. Ils ont manipulé les caractères pour rendre les informations locales ou globales plus saillantes, et ont ainsi réussi à montrer que l'effet de précédence globale pouvait être inversé si les informations locales étaient rendues suffisamment saillantes. Ces résultats seront plus détaillés au chapitre 3.

1.3.2.2 Mise en évidence expérimentale de la hiérarchie du traitement local/global de l'information en audition

En audition, quelques études ont été menées plus récemment pour étudier cette question de l'organisation du traitement local/global de l'information, avec l'équivalence forme/espace en vision pour hauteur/temps en audition (Bouvet et al., 2011 ; Justus and List, 2005 ; Ouimet et al., 2012 ; Sanders and Poeppel, 2007).

Justus and List (2005) ont d'abord proposé des mélodies de 9 notes successives dont la hauteur pouvait être modifiée à l'échelle d'une seule note (locale) ou d'un groupe de notes (globale). Ces stimuli ont été repris dans la suite, la tâche des participants étant de détecter les variations locales ou globales de la hauteur des notes. L'effet de précédence globale a ainsi pu être révélé dans la modalité auditive. Bouvet et al. (2011) ont montré que les participants étaient plus rapides et plus précis pour détecter les variations glo-

bales, c'est-à-dire pour traiter l'information à l'échelle globale, aussi bien dans la modalité auditive que dans la modalité visuelle. Ils ont également constaté une plus grande interférence des informations globales sur les informations locales que l'inverse. D'autres études Black et al. (2017) ; Ouimet et al. (2012) ont ensuite confirmé l'existence de cet effet de précédence globale dans le traitement de l'information auditive.

Finalement, Susini et al. (2020) ont retrouvé ces résultats dans un paradigme et un cadre d'analyse utilisant la Théorie de la Détection du Signal (TDS). Dans leur protocole, les notes ont en outre été regroupées temporellement en trois triplets, ce qui permet de mieux distinguer les échelles locale et globale. Leurs résultats, analysés dans le cadre de la TDS, rendaient compte des variations de sensibilité des participants pour détecter les variations à une échelle locale ou globale indépendamment des variations des stratégies de réponse des participants.

L'influence de la saillance des stimuli sur la hiérarchie du traitement local/global (voir section 1.3.2.1) n'a, quant à elle, pas été menée en audition. Ce sera l'objet du chapitre 3.

EN RÉSUMÉ

— La perception d'une scène sonore complexe met en jeu des principes d'analyse de scènes auditives, qui sont sous influence de mécanismes ascendants et descendants, notamment de processus attentionnels.

— En situation de multi-exposition, la perception d'une scène sonore dans sa globalité relativement à celle de chaque source suit des principes complexes impliquant notamment des influences réciproques.

— Cette influence de la perception de l'ensemble sur celle de ses parties, et réciproquement, est caractérisée par le principe de primauté du traitement holistique de l'information.

— Un traitement prioritaire de l'information à l'échelle globale a été mis en évidence dans la modalité visuelle dans des expériences utilisant des paradigmes sondant et comparant le traitement de l'information au niveau local et au niveau global. Cette hiérarchie peut être affectée (inversée) par la saillance des stimuli.

— La priorité du traitement global de l'information a été retrouvée dans la modalité auditive. L'influence de la saillance des stimuli n'y a pas encore été étudiée.

1.4 La perception de scènes sonores environnementales

Les préoccupations sociétales sur le bruit ont généralement porté sur la réduction du bruit dans l'environnement, et ont ainsi conduit à le voir comme un facteur ayant un impact négatif sur la santé et dont il faut réduire le niveau de nuisance (OMS, 2011 ; Trudeau et al., 2018). Or, la réduction du niveau de bruit n'est pas nécessairement liée à une meilleure qualité de vie, et l'expérience vécue lors de la perception d'un environnement sonore ne peut pas se limiter aux caractéristiques physiques du bruit perçu (Kang, 2006 ; Kang and Schulte-Fortkamp, 2018). Il peut être intéressant de considérer l'environnement sonore tel qu'il est perçu et vécu par un individu dans un certain contexte. C'est l'approche de la communauté scientifique qui s'intéresse à la notion de paysage sonore.

1.4.1 Le paysage sonore : cadre conceptuel

Introduit par Southworth (1967) puis développé par Schafer (1977), le paysage sonore consiste en "l'environnement acoustique tel que perçu et/ou vécu par un ou des individus, selon le contexte" selon l'organisation internationale de normalisation (ISO, 2014) [3].

Pour Schafer (1977), les composantes d'un environnement sonore doivent être considérées telles qu'elles sont perçues et avec le sens qu'elles peuvent revêtir pour les individus qui y sont immergés. La perception d'un paysage sonore se fait toujours par un individu, qui construit des représentations mentales et donne un sens aux objets de son environnement. Cette idée a ensuite été reprise et enrichie par des propositions de cadres conceptuels de plus en plus élaborés et prenant en compte de plus en plus de composantes impliquées dans la perception d'un paysage sonore.
Truax (1984) a par exemple enrichi ce principe en proposant un premier

3. traduit par l'auteur

cadre conceptuel dans lequel le son, en tant qu'entité physique, est présenté comme un médiateur entre l'auditeur, en tant qu'individu avec des mécanismes cognitifs internes et des associations de sens, et l'environnement. Kang and Schulte-Fortkamp (2018) ont, plus tard, proposé un cadre conceptuel fondé sur une idée similaire, et dans lequel le paysage sonore est présenté comme la construction mentale de l'environnement sonore par un individu. Herranz-Pascual et al. (2010) ont également proposé leur cadre, centré autour de trois composantes principales : l'individu, le contexte (l'environnement physique du lieu) et l'activité (le rapport de l'individu à l'environnement sonore, l'usage du lieu).
Les cadres proposés successivement ont ainsi cherché à positionner l'environnement sonore tel que perçu dans son contexte d'une manière de plus en plus détaillée.

Informée par ces différentes propositions de cadres conceptuels, l'organisation internationale de normalisation (ISO) a elle-même proposé un cadre théorique pour définir et contextualiser le paysage sonore. Il inclut différentes composantes, comme le contexte, les sources sonores ou la sensation auditive et leurs interactions. Plus précisément, le contexte inclut des sources sonores, qui produisent un environnement acoustique, qui donne lieu à des sensations auditives, et qui, une fois interprétées, mènent à une réponse de l'individu dans son contexte (ISO, 2014).

Les cadres conceptuels ont toujours visé à embrasser la complexité du paysage sonore et à dépasser la vision d'environnement sonore comme simples sommes d'informations physiques. Cependant, les besoins d'évaluation et de standardisation des mesures des environnements sonores (par les acteurs urbains - architectes, urbanistes - prenant en compte la dimension sonore dans le processus d'aménagement) se heurtent à cette vision complexe et protéiforme du paysage sonore.

1.4.2 Composantes principales du paysage sonore

Des travaux de recherche plus récents se sont ainsi concentrés sur l'exploration des composantes principales de la perception d'un paysage sonore, afin d'en harmoniser son évaluation.

Axelsson et al. (2010) ont demandé à des participants d'évaluer des paysages sonores selon une cinquantaine d'attributs (tels que calme, stimulant, vivant, irritant, ennuyeux, complexe, familier par exemple). Puis, une analyse en composantes principales (ACP) a permis de révéler deux composantes en particulier : *pleasantness* et *eventfullness* (que l'on peut traduire par *agrément* et *caractère animé*). Ainsi, tout attribut pour qualifier un paysage sonore peut être considéré comme une combinaison de ces deux composantes orthogonales. Par exemple, un environnement "vivant" est agréable et animé, un environnement "chaotique" est animé mais désagréable, et un environnement calme est agréable et peu animé.
D'autres chercheurs ont mis en place des méthodologies similaires, tels que Cain et al. (2013) qui se sont concentrés sur les émotions ressenties en présence de paysages sonores. Ils ont également mis en évidence deux composantes principales : *calmness* et *vibrancy* (que l'on peut traduire par *calme* et *caractère stimulant*).

Axelsson et al. (2012) ont dérivé de leurs travaux précédents un protocole de collecte de données pour l'évaluation des paysages sonores : le Swedish Soundscape Quality Protocol (SSQP). Cet outil a ensuite été intégré aux "échelles de qualité affective perçue" (Perceived Affective Quality Scales - PAQS) de la norme ISO relative aux paysages sonores (ISO, 2019). Ce protocole inclut 8 attributs : *pleasant, annoying, calm, chaotic, eventful, uneventful, vibrant*, et *monotonous*, présentés en figure 1.6. La traduction de ces termes fait l'objet de travaux visant à rendre ces protocoles utilisables dans différentes langues (Aletta et al., 2023 ; Tarlao et al., 2023). On les traduit généralement de la manière suivante en français : agréable, désagréable, calme, chaotique, animé, amorphe, stimulant et ennuyeux.

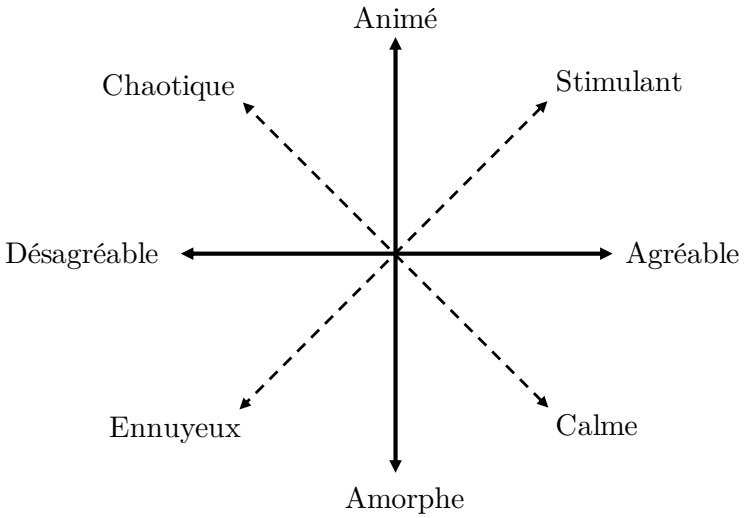

FIGURE 1.6 – Perceived Affective Quality Scales (PAQS) tel que présenté dans la norme ISO (2019)

Ainsi, des modèles dimensionnels ont permis de mettre en évidence les composantes principales de la perception des paysages sonores, réduisant ainsi la caractérisation complexe de leur perception à une évaluation de quelques attributs perceptifs. Aujourd'hui, cette évaluation se fait selon un protocole suggéré par une norme ISO (ISO, 2019) qui tente de l'uniformiser et d'en préciser les modalités : mesures des attributs perceptifs selon des échelles de Likert ou des échelles sémantiques différentielles, par exemple. Malgré ces efforts d'uniformisation, un consensus scientifique sur les échelles d'évaluation de l'environnement sonore n'est aujourd'hui pas complètement établi.

1.4.3 L'évaluation du paysage sonore

Indicateurs de la qualité de l'environnement sonore

Évaluer un environnement sonore demande, entre autres, de pouvoir situer cet environnement dans l'espace des dimensions perceptives présenté en figure 1.6. Or, un outil d'évaluation de la qualité d'un paysage sonore se doit d'être pratique et simple à mettre en oeuvre. La collecte de données sur le terrain auprès de personnes n'étant pas systématiquement réalisable, il peut être bénéfique de pouvoir estimer la qualité d'un environnement sonore donné à l'aide d'indicateurs psychoacoustiques qui se rapprochent des attributs perceptifs présentés précédemment.

La caractérisation de la qualité des environnements sonores par des indicateurs acoustiques ou psychoacoustiques a déjà fait l'objet de diverses propositions. Des paramètres physiques comme le niveau sonore équivalent L_{eq} et ses dérivés peuvent être partiellement corrélés au désagrément sonore (Aumond et al., 2017a). Des expériences *in situ* associant évaluation de l'environnement et mesures acoustiques ont été menées pour relever des données d'évaluation de la qualité de l'environnement sonore en parallèle de données acoustiques (Nilsson et al., 2007 ; Ricciardi et al., 2015). Les indicateurs dérivés des données de niveau sonore (niveau de pression acoustique, sonie, niveau moyenné sur les bandes de tiers d'octaves) y expliquaient une part de la variabilité des données d'évaluation de la qualité de l'environnement sonore (Gozalo et al., 2015). En France, deux observatoires du bruit, Bruitparif et Acoucité, ont proposé un indice (harmonica) pour prendre en compte à la fois le niveau énergétique du bruit de fond et celui des bruits émergents (Mietlicki et al., 2014) et essayer de caractériser les impressions du grand public en situation d'exposition au bruit urbain.

Cependant, ces études ont également pointé du doigt l'importance de la prise en compte des sources sonores en particulier. Les modèles proposés par Ricciardi et al. (2015) qui accordent plus d'importance aux sources émer-

gentes sont plus proches des données d'évaluation des participants, et Nilsson et al. (2007) soulignent l'importance du "développement d'indicateurs acoustiques de l'audibilité des sources dans les paysages sonores". Ces résultats sont cohérents avec ceux d'autres études qui ont montré que les indicateurs sonores énergétiques ne suffisent pas pour embrasser la complexité d'un paysage sonore et donc la perception qu'en ont les individus (Aletta et al., 2016 ; Berglund et al., 2002 ; Waye and Öhrström, 2002).

L'importance de la considération des sources sonores

L'influence des sources sonores sur la qualité perçue des paysages sonores, et leurs caractéristiques, semble ainsi importante dans la perception d'un environnement. Leur nature, par exemple, peut affecter la perception de l'agrément sonore. Lavandier and Defréville (2006) ont ainsi montré que des sources sonores naturelles (chants d'oiseaux) contribuaient à rendre en laboratoire des environnements sonores urbains plus agréables, des sources mécaniques (passages de véhicules) plus désagréables. Guastavino (2006) a retrouvé ce résultat par une analyse sémantique de réponses à des questionnaires : les sources sonores naturelles étaient associées à des jugements positifs, les sources mécaniques à des jugements négatifs. Dans des questionnaires *in situ*, Nilsson et al. (2007) ont enfin confirmé ce résultat en relevant des corrélations entre qualité de l'environnement sonore perçu et identification de sources sonores : corrélation négative pour les sources technologiques, corrélation positive pour les sources naturelles. Il y a ainsi aujourd'hui un consensus sur l'effet de la nature des sources sonores en général, les sources naturelles ayant un effet positif sur la qualité perçue de l'environnement sonore, les bruits de circulation ayant un effet négatif, et les sons d'origine anthropique (des voix par exemple) pouvant avoir un effet positif ou négatif (Axelsson et al., 2012 ; Nilsson and Berglund, 2006 ; Yang and Kang, 2005).

Il semble ainsi avisé, pour prédire l'agrément sonore d'un environnement, de prendre en compte des paramètres liés à ces sources, comme par exemple leur temps de présence ou leur fréquence d'occurrence dans une scène (La-

vandier et al., 2021 ; Lavandier and Defréville, 2006).
De plus, la composition des scènes sonores est régie par l'organisation temporelle de ces sources. Des études ont ainsi montré que l'agrément sonore était mieux prédit si l'on incluait des paramètres relatifs aux variations temporelles du son, et à la contribution de sources sonores spécifiques (Bockstael et al., 2011 ; Ricciardi et al., 2015). Ces prédictions peuvent être améliorées en introduisant d'autres indicateurs, la variabilité du niveau dans le temps par exemple (Nilsson and Berglund, 2006).

Ainsi, l'évaluation du paysage sonore passe par la mise en évidence d'indicateurs visant à en caractériser les attributs perceptifs. Une association partielle entre la qualité de l'environnement sonore et des paramètres acoustiques fondés sur les niveaux d'énergie acoustique moyens peut être observée. Cette association reste limitée, et il est nécessaire de déterminer des indicateurs se rapprochant plus encore de la perception des environnements sonores, notamment en prenant en compte les sources sonores perçues comme émergentes.

1.4.4 Saillance et paysage sonore

L'environnement sonore est rarement le centre de l'attention volontaire d'un individu (Botteldooren et al., 2015). Au-delà des échanges verbaux auxquels ils peuvent participer, les usagers de l'espace public n'ont en effet que peu d'intérêt à prêter de l'attention aux sons qui les entourent, un chant d'oiseau par exemple. Dès lors, la plupart des sollicitations environnementales ne franchissent pas leur filtre attentionnel. Elles peuvent néanmoins avoir des effets sur les individus (OMS, 2011) : elles contribuent par exemple au stress ressenti (Gatersleben and Griffin, 2017) et à la gêne (Berglund, 1998), avec un effet bénéfique de la présence de la nature pour réduire le stress et les réponses physiologiques associées (Bogdanov et al., 2022), et réduire la gêne exprimée (Van Renterghem and Botteldooren, 2016).

Les sollicitations ignorées, par définition, ne contribuent pas à l'évaluation et l'appréciation consciente de l'environnement sonore (Filipan et al., 2019). Ce n'est que lorsqu'un bruit attire l'attention qu'il peut en effet être appréhendé comme une source sonore (Botteldooren et al., 2015). Il suffit alors d'une faible quantité d'évènements sonores saillants qui "marquent" l'environnement pour pouvoir lui donner une identité (Oldoni et al., 2015). Son évaluation devient alors possible. C'est pourquoi l'identification des facteurs pouvant accroître la saillance d'une source sonore dans son environnement est un aspect important de la perception des paysages sonores.

Les caractéristiques acoustiques font partie des principaux facteurs qui peuvent influencer la saillance d'une source sonore. C'est d'ailleurs sur cette idée que s'est fondé le fonctionnement de modèles de saillance actuels (voir section 1.2.3). Mais la congruence et la familiarité de la source dans son environnement peuvent également être déterminantes : les sons incongrus ou familiers sont en effet plus susceptibles d'être remarqués (Gygi and Shafiro, 2011 ; Kirmse et al., 2009).

La saillance des différents sons que constituent un paysage sonore semble ainsi prépondérante dans sa perception et son appréciation. L'évaluation de l'environnement sonore semble en effet principalement liée à celle des sources saillantes, qui peuvent émerger à cause de leurs caractéristiques acoustiques.

EN RÉSUMÉ

— La perception d'un environnement se fait toujours par un individu, dans un contexte donné : le concept de paysage sonore restitue l'environnement sonore dans un cadre mettant en jeu des composantes plus complexes que la seule perception d'informations acoustiques.

— La perception d'un paysage sonore peut néanmoins être synthétisée en évaluant l'environnement sonore sur des dimensions qui caractérisent l'espace perceptif des paysages sonores. Les caractères agréable/désagréable et animé/amorphe sont notamment les deux dimensions majeures du paysage sonore.

— Évaluer un environnement sonore, entre autres son caractère agréable, implique non seulement de considérer des indicateurs psychoacoustiques propres au signal physique, mais également de considérer le rôle des sources qui le constituent spécifiquement.

— Cette importance des sources sonores dans l'évaluation du paysage sonore invite à s'intéresser précisément à la saillance, comme potentiel déterminant de la perception et de l'appréciation des scènes environnementales.

1.5 Objectifs et organisation de la thèse

Ainsi, les préoccupations sociétales remontées aux pouvoirs publics ces dernières années font part d'une préoccupation grandissante concernant les sources de bruit émergentes dans l'environnement (cf. Introduction). Celles-ci mettent en jeu toute la "complexité des interactions entre les divers paramètres physiques, physiologiques, humains et cognitifs impliqués dans les relations bruit-santé" (ANSES, 2013). Or, nous avons vu dans ce chapitre que la notion de saillance est justement au coeur de ces interactions entre paramètres physiques et traitement cognitif des informations sonores par le sujet humain (cf. section 1.2). Dès lors, c'est bien sur cette notion de saillance auditive que nous souhaitons porter nos efforts de recherche.

Les avancées dans les travaux de modélisation ont permis d'aboutir à des modèles capables de prédire la saillance dans des scènes sonores environnementales (cf. section 1.2.3). Ces modèles se fondent entre autres sur des hypothèses concernant les relations entre les paramètres physiques d'une scène sonore et la saillance. Or, il n'y a pas eu à ce jour de mise en évidence expérimentale de la relation entre les variations des propriétés d'un son et sa saillance (autrement dit la composante *stimulus-driven* de l'attention, cf. 1.2.5). Peut-on mettre en évidence la relation entre la capture attentionnelle produite par un son et les propriétés sonores de ce dernier ? Et quelles sont les propriétés de cette relation ? Ces questions seront traitées au chapitre 2.

On sait de plus que la perception et l'analyse d'une scène auditive peuvent mettre en jeu des mécanismes complexes (cf. section 1.3). Nous avons notamment vu que cette analyse est menée sous influence de mécanismes ascendants et descendants, notamment attentionnels, et dont les influences relatives et mutuelles ne sont pas encore connues. Or, l'effet de primauté du traitement holistique semble être affecté par la saillance dans la modalité visuelle. Qu'en est-il dans la modalité auditive ? La saillance engendrée par des variations de timbre étudiées au chapitre 2 a-t-elle un effet sur ce phénomène de primauté du traitement holistique ? Nous apporterons des réponses à ces interrogations au chapitre 3.

Enfin, il apparaît que la perception et l'appréciation de paysages sonores mettent en jeu une variété de mécanismes complexes liés à l'individu, son activité, ou le contexte dans lequel il évolue (cf. section 1.4). L'évaluation de la qualité de ces paysages sonores se fait par le biais d'attributs perceptifs qui en caractérisent les dimensions principales, entre autres la dimension agréable/désagréable. Or, la prédiction de l'agrément semble devoir dépasser l'utilisation d'indicateurs acoustiques et prendre en compte la nature et l'effet des sources qui composent une scène sonore. Dès lors, peut-on mettre en évidence le rôle de la saillance auditive sur la perception du désagrément dans des scènes sonores environnementales ? Nous nous intéresserons à ce point au chapitre 4.

Les questions posées et les études menées pour y répondre dans le cadre de cette thèse suivent une progression dans le sens d'une complexité croissante. D'abord étudiée dans un paradigme mettant en jeu des séquences sonores simples et contrôlées (chapitre 2), la notion de saillance y est ensuite explorée dans le cadre de séquences suivant une organisation temporelle plus complexe et mettant en jeu une hiérarchie dans le traitement de l'information auditive (chapitre 3), puis au sein de scènes sonores environnementales, séquences complexes et non-contrôlées par excellence (chapitre 4).

Chapitre 2

Modulation de la capture attentionnelle par des attributs du timbre

“Le soleil brille pour tout le monde.”

Proverbe français

Ce chapitre explore la manière dont les propriétés physiques d'un son, par les sensations auditives qu'elles suscitent, peuvent moduler sa saillance, donc sa capacité à capturer l'attention d'un auditeur indépendamment de sa volonté. Cette étude est motivée par les arguments exposés en section 1.3 : s'il existe des pistes de déterminants de la saillance auditive, dont les modèles de saillance relèvent les évolutions temporelles, et que la composante bottom-up de l'attention, dite *stimulus-driven*, est reconnue sur un plan théorique, il n'y a pas encore eu de mise en évidence expérimentale, à notre connaissance, d'une modulation de l'attention par les variations de ces paramètres.

Pour lever ce verrou, il nous fallait en premier lieu mettre en place une mesure de capture attentionnelle dans laquelle il était possible de manipuler les paramètres sonores de notre choix. Ce point a fait l'objet d'une revue de bibliographie visant à recenser les différents paradigmes expérimentaux pouvant permettre de mesurer un phénomène de capture attentionnelle. Cette revue fait l'objet de la première partie de ce chapitre. Nous présentons ensuite, dans une deuxième partie, les expériences nous ayant permis d'évaluer et d'adapter un paradigme à nos besoins. Enfin, nous détaillons comment nous avons mis en évidence la modulation bottom-up de l'attention par des attributs du timbre, à partir de ce paradigme. Cette dernière partie est présentée sous la forme de l'article de revue publié durant la thèse (Bouvier et al., 2023b).

2.1 Mesure de la capture attentionnelle

La saillance, du fait de sa nature, n'est pas une grandeur que l'on peut mesurer directement. En effet, nous l'avons définie comme la composante bottom-up de l'attention, c'est-à-dire la capacité d'un stimulus à capturer l'attention d'un auditeur indépendamment de sa volonté. Pour pouvoir la mesurer, il faut donc parvenir à observer à quel point l'attention d'un individu est capturée par un stimulus donné, sans que celui ne l'ait souhaité, c'est-à-dire sans que celui ne puisse prêter volontairement attention à ce stimulus. Les seules possibilités de mesure sont donc des mesures indirectes. Concrètement, cela implique d'avoir une mesure de l'attention d'un individu, et une manière

d'orienter son attention sur autre chose que l'objet dont on souhaite mesurer la saillance.

L'observation d'un phénomène de capture attentionnelle et sa mise en évidence ont d'abord été menées dans la modalité visuelle. Nous recensons donc ici les différentes méthodes expérimentales issues de la vision et qui pourraient être adaptées dans la modalité auditive, puis les méthodes existant déjà dans la modalité auditive.

2.1.1 Mesure de capture attentionnelle dans la modalité visuelle

Le phénomène de capture attentionnelle peut être mis en évidence de manière explicite ou de manière implicite (pour une revue, voir Simons (2000)).

On parle de capture explicite quand des participants prennent conscience d'un stimulus saillant et non pertinent pour réaliser une tâche qui les occupent. Plus précisément, les participants rapportent la présence ou non d'un stimulus non pertinent une fois la tâche réalisée. Le stimulus dont on souhaite mesurer la saillance doit nécessairement être non-pertinent. En effet, pour mesurer uniquement la composante "bottom-up" de l'attention, il faut que le participant ne recherche pas volontairement ce stimulus.
On parle de capture implicite quand on observe à quel point des participants peuvent ignorer des stimuli qu'ils savent non pertinents pour réaliser une tâche. On mesure la capacité du stimulus à capturer l'attention des participants en mesurant à quel point la performance dans la tâche est dégradée en sa présence.

La capture attentionnelle explicite se traduit ainsi par une prise de conscience du stimulus non pertinent, la capture implicite par une dégradation de performance sur la tâche à réaliser.

2.1.1.1 Capture attentionnelle visuelle explicite

Des études de capture attentionnelle explicite (Mack and Rock, 1998b ; Most et al., 2000 ; Neisser and Becklen, 1975 ; Newby and Rock, 1998 ; Simons and Chabris, 1999) ont montré que lorsque des participants sont concentrés sur un objet ou un événement en particulier dans une scène, ils peuvent ne pas remarquer certains autres objets pourtant évidents quand ils observent la scène dans sa globalité sans se concentrer sur une zone spécifique : on appelle cela, dans le domaine visuel, le phénomène de cécité attentionnelle. Un exemple populaire, le gorille apparaissant au milieu d'échanges de ballons (Simons and Chabris, 1999), a déjà été présenté en 1.2.4.

Neisser and Becklen (1975) ont montré en premier que lorsque des participants visionnent deux séquences vidéos superposées (deux vidéos distinctes dont les images sont superposées) et prêtent attention à une des deux séquences en particulier, ils ne remarquent quasiment jamais les évènements inattendus qui se produisent dans l'autre séquence. Mack and Rock (1998b) ont révélé l'existence d'une zone attentionnelle en dehors de laquelle nous sommes aveugles aux stimuli inattendus qui peuvent y apparaître brièvement. Dans leur expérience, les participants observaient le centre d'une croix. Lorsqu'un stimulus non attendu apparaissait dans un des quadrants de la croix, il avait plus de chances d'être détecté que lorsqu'il apparaissait en dehors de cette zone. Autrement dit, le niveau de cécité attentionnelle augmente lorsqu'on sort de la zone de focalisation de l'attention. Newby and Rock (1998) ont approfondi cette étude en évaluant le niveau de cécité attentionnelle en fonction de la distance du stimulus inattendu au centre de la zone d'attention et non pas de la zone de fixation du regard (en plaçant le centre de la croix à une certaine distance du point de fixation du regard). Ils ont montré que c'est la distance au centre de la zone d'attention qui importe, et que plus cette distance est grande, plus il y a de chances que se produise le phénomène de cécité attentionnelle. Most et al. (2000) ont également étudié l'influence de la distance du stimulus à la zone de focalisation de l'attention. Ils ont observé que plus l'objet inattendu était loin de cette zone sur laquelle l'attention des participants était portée, moins cet objet avait de chances d'être remarqué, et

ce même pour des stimuli visibles pendant une longue durée.

La capture attentionnelle explicite implique ainsi d'aller jusqu'à la prise de conscience du stimulus inattendu. C'est la conception la plus intuitive de ce phénomène : on sait qu'un stimulus a capté notre attention lorsqu'on en a pris conscience. Cependant, des méthodes permettent de mettre en évidence ce phénomène de capture attentionnelle sans nécessairement aller jusqu'à la prise de conscience. On parle alors de capture attentionnelle implicite.

2.1.1.2 Capture attentionnelle visuelle implicite

La capture attentionnelle implicite se produit lorsque la présence d'un stimulus non pertinent saillant se traduit par la dégradation de la performance des participants sur la réalisation d'une autre tâche. L'observation de cette dégradation de performance constitue la mesure implicite du phénomène de capture attentionnelle. Le cadre dans lequel ce phénomène est étudié est par exemple celui d'une tâche de recherche d'une cible parmi un ensemble d'items.

Mode de recherche visuelle et état attentionnel

Dans le cas de la recherche dans un ensemble d'items, on appelle "cible" l'objet particulier recherché par les participants, et "distracteurs" les autres items qui sont tous identiques. "Item" est le nom donné à n'importe quel objet, qu'il soit cible ou distracteur. Dans la modalité visuelle, on observe l'existence de deux modes de recherche distincts lorsqu'il s'agit de trouver une cible dans un ensemble de distracteurs (Maquestiaux, 2017) :

— un mode de recherche sérielle : l'individu parcourt les items un à un jusqu'à ce qu'il tombe sur la cible. Ce mode de recherche est adopté quand la cible est difficilement séparable des distracteurs.

— un mode de recherche parallèle : l'individu prend une photo de l'ensemble des items et repère la cible dans l'ensemble global. Ce mode de recherche est adopté quand la cible se démarque bien des distracteurs.

L'adoption de ces deux modes de recherche rappelle les étapes de la théorie d'intégration des caractéristiques de Treisman and Gelade (1980) (voir partie 1.1.3.4) : une première étape d'enregistrement parallèle des caractéristiques de la scène qui correspond au stade pré-attentionnel (mode de recherche parallèle), puis une deuxième étape d'identification des objets un à un nécessitant plus de ressources en attention (mode de recherche sérielle).

La fenêtre attentionnelle est de taille variable et dépend de l'état attentionnel du participant, qui dépend de la tâche que ce dernier doit réaliser. Ceci a été montré par des études qui examinaient les réponses de participants (temps de réaction : LaBerge (1983)) ou des mesures physiologiques (pupillométrie : Tkacz-Domb and Yeshurun (2018)). Elle correspond à la zone qui est mise en évidence dans les études mentionnées en section 2.1.1.1 (Mack and Rock, 1998b ; Most et al., 2000 ; Newby and Rock, 1998). Le mode de recherche parallèle implique une fenêtre attentionnelle large qui inclut l'ensemble des items présentés. Le mode de recherche sérielle implique une fenêtre attentionnelle plus réduite que l'individu déplace d'item en item.

La capture attentionnelle est modulée par la taille de cette fenêtre attentionnelle. Ainsi, Belopolsky et al. (2007) ont montré que les participants placés dans un état d'attention diffuse (fenêtre attentionnelle large) étaient plus facilement sujets à la capture attentionnelle par un singleton de couleur différente que quand ils étaient dans un état d'attention restreinte (fenêtre attentionnelle réduite). Ils ont affirmé que bien que la taille de cette fenêtre soit sous contrôle (le participant élargit volontairement sa fenêtre plus ou moins pour réaliser une tâche), un singleton saillant peut capturer l'attention par des mécanismes bottom-up. Ce phénomène de capture attentionnelle ne peut cependant avoir lieu que si le stimulus non pertinent censé capturer l'attention des participants survient à l'intérieur de leur fenêtre attentionnelle (Belopolsky and Theeuwes, 2010). Dans leur expérience, Belopolsky and Theeuwes (2010) ont montré que lorsque cette fenêtre attentionnelle était réduite (en focalisant l'attention des participants sur une présentation sérielle rapide au centre de l'écran par exemple) et que le stimulus non pertinent appa-

raissait en dehors de la fenêtre, le phénomène de capture ne se produisait plus.

Paradigme du singleton additionnel

Dans ce paradigme, les participants doivent retrouver une cible parmi plusieurs autres items tous identiques, les distracteurs (Theeuwes, 1992,9). Cependant, un nouvel item peut être rendu différent des distracteurs. On l'appelle le "singleton". Ce singleton n'a pas de rapport avec la tâche de recherche : il est dit "non pertinent", c'est-à-dire que lui prêter attention ne permet pas d'être plus performant dans la tâche de recherche. Ce singleton n'est jamais la cible. Les participants doivent donc l'ignorer.

Le temps de réponse est comparé dans deux conditions. Dans la première condition, la cible est seule en présence des distracteurs. Dans la seconde, la cible est en présence des distracteurs et du singleton. Le phénomène de capture attentionnelle est mis en évidence par un allongement des temps de réponse dans la condition où le singleton non pertinent est présent.

Une des premières études, celle de Pashler (1988), a montré que la recherche d'un item qui se différenciait des distracteurs par sa forme était ralentie quand un autre item se différenciait par sa couleur. Il a montré que ce ralentissement n'était vérifié que dans le cas où c'était un singleton qui se démarquait par la dimension non pertinente (la couleur ici). Des variations de couleur sur des sous-ensembles des distracteurs (les distracteurs avaient une couleur différente en fonction de leur appartenance à une bande verticale ou horizontale du panneau de présentation par exemple) ne donnaient pas lieu à ce phénomène.

Theeuwes (1992) a présenté à des participants un certain nombre de losanges verts, parmi lesquels se trouvait un rond vert. Dans chaque item, un segment était présenté avec une orientation aléatoire (plusieurs inclinaisons possibles tous les 22,5°). Il était demandé aux participants de dire si le segment présent dans le rond vert était horizontal ou vertical. Dans la moitié

des essais, un des losanges était rouge. Les participants étaient informés du fait que le losange rouge ne contenait jamais la cible et donc qu'il était non pertinent. La présence de ce singleton se traduisait par un allongement des temps de réponse, prouvant que ce singleton, rendu unique sur une dimension non pertinente pour la réalisation de la tâche, capturait l'attention. Il a conclu que dans le cadre de cette tâche (de recherche de cible unique), les stimuli capturent l'attention des participants en fonction de leur saillance. Autrement dit, si le singleton additionnel est saillant, un effet de sa présence est observé sur les temps de réponse.

Il est important de noter l'importance de la tâche à réaliser dans le cadre de ce paradigme. La recherche d'une cible rendue unique place les participants dans une stratégie de recherche de singleton (Pashler, 1988 ; Theeuwes, 1992 ; Yantis, 2000). Cette recherche d'un élément différent et unique peut donc être ralentie dans le cas où un autre singleton, le singleton additionnel, est présent.

Paradigme de caractéristique non pertinente

Dans ce paradigme, les participants recherchent une cible et un des items peut être rendu différent par une caractéristique en particulier (voir figure 2.1). Ce dernier peut être l'un des distracteurs ou la cible elle-même, c'est-à-dire que l'item qui se détache par une propriété particulière peut être le même que celui que les participants recherchent. En condition contrôle, il n'y a pas de caractéristique particulière modifiée : on a donc simplement la cible parmi les distracteurs. Si la tâche est suffisamment compliquée, le temps de recherche augmente avec le nombre d'items présentés (Yantis and Hillstrom, 1994). Lorsque c'est la cible qui est rendue différente par une caractéristique particulière qui la rend saillante, le temps de réponse augmente moins rapidement avec le nombre d'items présents. L'item rendu distinctif par la caractéristique particulière devrait en effet être le premier considéré, et donc s'il s'agit de la cible, elle est trouvée à la même vitesse quel que soit le nombre d'items présents. La capture attentionnelle est mise en évidence par une diminution de la pente du temps de réponse en fonction du nombre d'items présentés

dans la condition où la cible a une caractéristique particulière (voir figure 2.2).

Les expériences réalisées avec ce paradigme ont montré que les apparitions abruptes d'items capturent l'attention. Yantis and Jonides (1984) ont en effet fait apparaître des lettres derrière des figures de 8 en bâtons (voir figure 2.1) : un écran avec un certain nombre de 8 était présenté aux participants, puis certains segments des 8 s'effaçaient (en 80 ms) pour laisser apparaître les lettres. Une lettre pouvait être présentée de manière abrupte quand elle apparaissait à un endroit où il n'y avait pas de 8 avant.

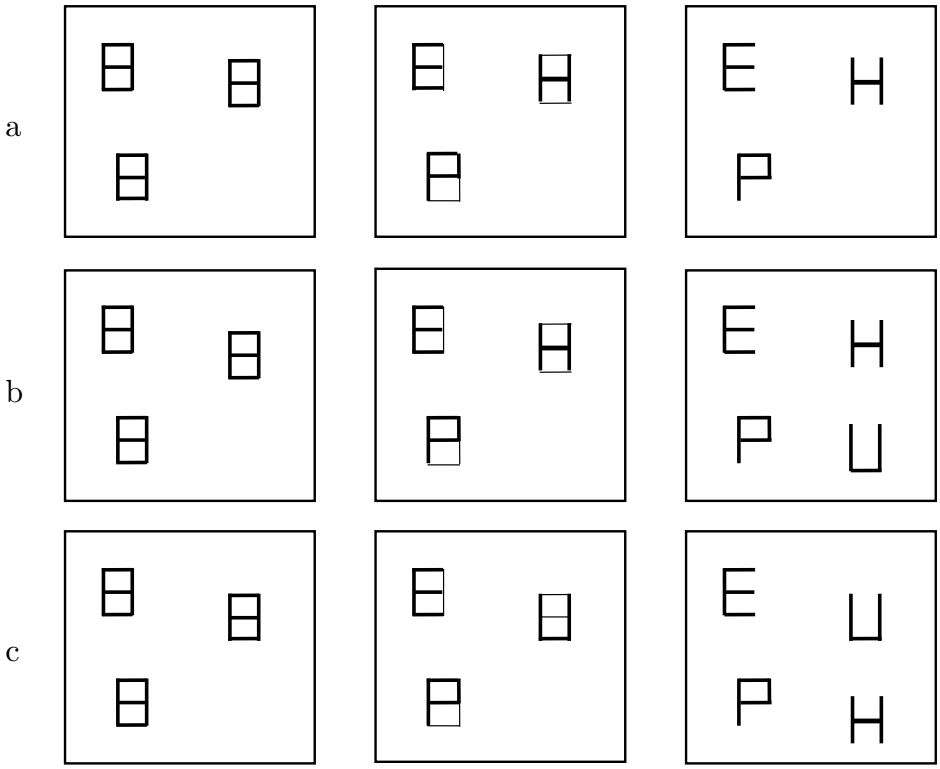

FIGURE 2.1 – Exemple générique illustrant le paradigme de caractéristique non pertinente. Ici la lettre H est la cible. Les panneaux sont présentés successivement dans l'ordre de gauche à droite. A gauche, le panneau initial avec les figures en 8 annonçant l'emplacement futur des lettres, au centre certains segments sur les 8 commencent à s'estomper, à droite les lettres sont pleinement apparues. Condition a : condition contrôle, pas d'apparition abrupte. Condition b : la cible est apparue progressivement, un distracteur est apparu abruptement. Condition c : la cible est apparue de manière abrupte.

Il a été observé que, lorsque la lettre qui apparaissait de manière abrupte était la cible (condition c sur les figures 2.1 et 2.2), la pente des temps de réponse en fonction du nombre d'items était plus faible que dans le cas où il n'y avait pas de présentation abrupte (condition a). De plus, la pente des temps de réaction, dans le cas où il n'y avait pas de présentation abrupte, était plus faible que dans le cas où c'était une lettre qui n'était pas la cible qui apparaissait de manière abrupte (condition b).

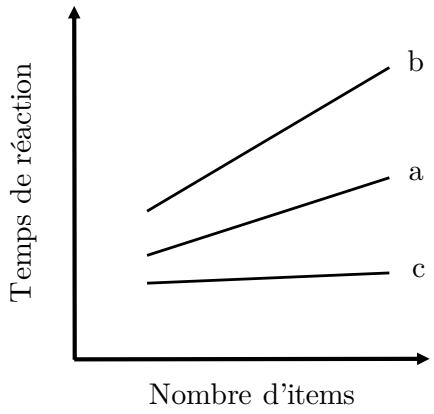

FIGURE 2.2 – Allure schématique des résultats observés dans le cadre du paradigme de caractéristique non pertinente. Quand le nombre de distracteurs augmente, les temps de réaction pour détecter la lettre cherchée augmentent. En condition c (apparition abrupte de la cible), le temps de réaction augment moins rapidement qu'en condition a (condition contrôle). En condition b (apparition abrupte d'un distracteur), le temps de réaction augmente plus rapidement qu'en condition contrôle.

Remington et al. (1992) ont mis en évidence l'incapacité à ignorer les apparitions abruptes, car même lorsque les participants étaient informés du fait qu'il fallait ignorer ces apparitions soudaines (un flash en amont sur une des zones de présentation des items à venir dans le cas de leur expérience), les temps de réaction pour la détection de la cible étaient allongés.

La tâche à réaliser dans ce paradigme ne semble en revanche pas permettre de tester d'autres caractéristiques (couleur, luminosité...). En effet, Jonides and Yantis (1988) ont montré que des différences de couleur ou d'intensité

ne pouvaient pas créer de capture attentionnelle : dans les conditions où la cible était de couleur unique ou de brillance différente, les temps de réaction augmentaient avec le nombre d'items présentés, comme dans la condition où la cible n'était pas rendue différente. Seules les apparitions soudaines donnaient lieu à une capture attentionnelle. Hillstrom and Yantis (1994) ont testé différents paramètres qui mettaient les items en mouvement : oscillation, scintillement, texture qui varie, points qui tournent autour de l'item... Leurs résultats ont montré que les temps de réaction augmentent avec le nombre d'items présentés, et ceci même quand c'est la cible qui est rendue mouvante. Il n'y a donc pas d'effet du mouvement des items, seules les apparitions soudaines capturent l'attention dans ce cadre expérimental.

Ce paradigme permet de mettre en évidence un phénomène de capture attentionnelle, mais la tâche demandée et l'état attentionnel des participants correspondant à cette tâche semblent ne donner lieu à ce phénomène que pour des items apparaissant de manière abrupte. Or, dans le cadre du paradigme du singleton additionnel, des variations de couleur donnent bien lieu à un phénomène de capture attentionnelle. Il semblerait donc que ce phénomène soit contingent de l'état attentionnel mis en place pour pouvoir réaliser la tâche.
Cette hypothèse qui apparaît au regard des paradigmes détaillés précédemment a été testée à l'aide d'un quatrième paradigme qui utilise des techniques d'amorçage.

Paradigme d'amorçage

Dans ce paradigme, la recherche visuelle est guidée par une amorce : juste avant la présentation des items, des indicateurs sont positionnés à l'emplacement futur des items (Folk and Remington, 1998 ; Folk et al., 1992,9 ; Gibson and Kelsey, 1998 ; Warner et al., 1990). On parle d'amorçage. Les participants sont informés du fait que l'amorçage est non pertinent : il n'est pas utile pour être meilleur dans la réalisation de la tâche. L'amorçage peut être rendu valide quand l'indicateur qui est situé là où sera révélée la cible est

rendu particulier (par sa couleur par exemple). Il peut être invalide lorsque l'indicateur rendu particulier n'est pas à la place de la future cible. Il peut être neutre (condition contrôle) quand aucun indicateur n'est rendu particulier (voir figure 2.3).

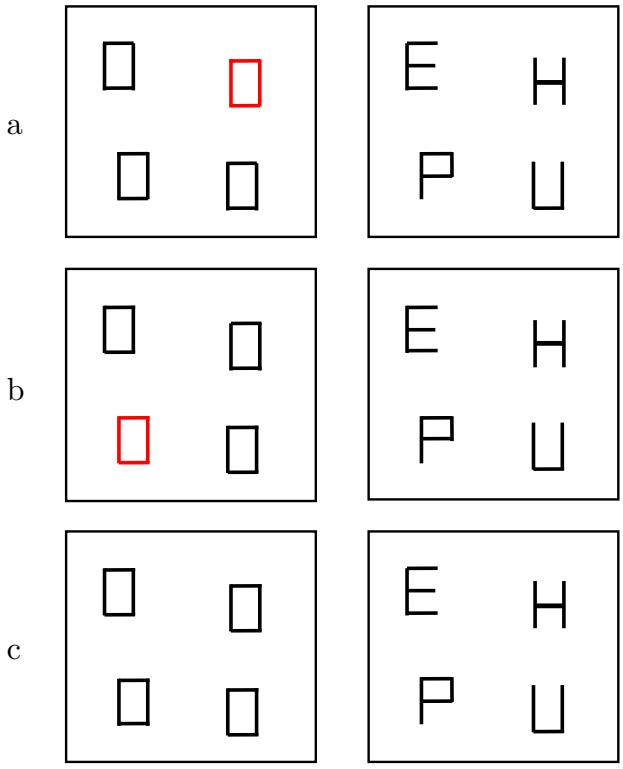

FIGURE 2.3 – Exemple générique illustrant le paradigme d'amorçage. Ici encore, la lettre H est la cible. Les panneaux sont également présentés successivement dans l'ordre de gauche à droite. A gauche, le panneau initial avec les indicateurs annonçant l'emplacement futur des lettres, à droite les lettres sont apparues. Condition a : condition d'amorçage valide. Condition b : condition d'amorçage invalide. Condition c : amorçage neutre, condition contrôle

L'effet de capture attentionnelle est mis en évidence quand les temps de réponse sont plus courts dans la condition d'amorçage valide (condition a sur les figures 2.3 et 2.4), que dans la condition contrôle (condition c), et plus longs dans la condition d'amorçage non-valide (condition b) que dans la condition contrôle.

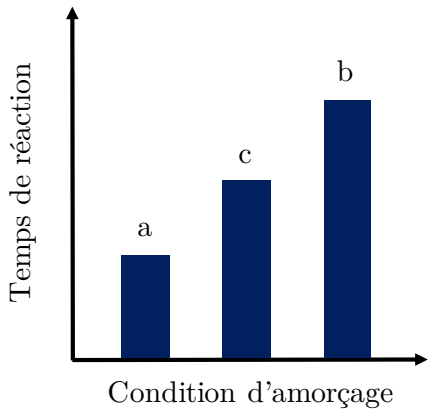

FIGURE 2.4 – Allure schématique des résultats observés dans le cadre du paradigme d'amorçage. Les temps de réaction dépendent de la condition d'amorçage. Dans l'ordre du plus court au plus long : amorçage valide (a), condition contrôle (c), amorçage non valide (b).

Folk et al. (1992) ont testé différentes combinaisons de paramètres permettant de marquer l'amorçage d'une part et la cible d'autre part. Ils ont ainsi montré que si l'amorçage et la cible se démarquaient par une variation sur une même caractéristique (la couleur ou la soudaineté de l'apparition), on observait un phénomène de capture attentionnelle. En revanche, ce phénomène n'était pas mis en évidence quand l'amorçage et la cible se démarquaient grâce à deux caractéristiques différentes. Par exemple, si un amorçage invalide était marqué par la soudaineté de l'apparition de l'indice de position, on observait des temps de réaction plus grands pour trouver une cible qui se démarquait par son apparition également, mais pas pour une cible qui se démarquait par sa couleur.

Folk et al. (1994) ont confirmé cette contingence entre l'état attentionnel des participants et la capture attentionnelle : les amorçages par le mouvement (rotation, translation) ont permis d'observer un effet quand la cible était démarquée par le mouvement également, mais pas quand elle était démarquée par la couleur. Ainsi, si l'état attentionnel est réglé pour détecter les stimuli dynamiques (mouvement, apparition soudaine) - respectivement statiques (couleur, brillance), seuls les stimuli dynamiques - respectivement statiques - capturent l'attention. De plus, Folk et al. (1994) ont remarqué une asymétrie

dans l'interaction de certaines caractéristiques permettant l'amorçage et la distinction de la cible, au sein des caractéristiques de mouvement (qui sont donc bien compatibles pour produire un effet). Un amorçage par le mouvement a un effet sur la détection de cibles qui se distinguent par le mouvement, mais pas par la soudaineté d'apparition. À l'inverse, un amorçage par la soudaineté d'apparition a un effet pour les cibles qui se distinguent par n'importe laquelle des deux caractéristiques. Cette asymétrie est attribuée d'après eux à une différence de saillance entre les stimuli en mouvement (moins saillants) et les stimuli qui apparaissent soudainement (plus saillants). Folk et al. (1994) sont en effet ensuite parvenus, en rendant les stimuli en mouvement plus saillants, à obtenir le même effet quelle que soit la caractéristique qui définissait l'amorçage et la cible. Ils ont donc conclu qu'une fois l'état attentionnel réglé pour pouvoir observer l'effet de capture attentionnelle, celui-ci est dépendant de la saillance des stimuli utilisés.

Gibson and Kelsey (1998) ont même montré qu'il y a une contingence entre la capture attentionnelle et les caractéristiques du panneau de présentation des items dans lequel il faut trouver la cible (et pas seulement la caractéristique qui démarque seulement la cible). Dans leur expérience, un amorçage par la couleur rouge a permis de mettre en évidence une capture attentionnelle si toutes les lettres (parmi lesquelles se trouve la cible) étaient ensuite présentées en rouge, pas si elles étaient présentées en blanc. Le mode de présentation conditionne ainsi l'état attentionnel des participants, et donc la possibilité d'observer le phénomène de capture attentionnelle.

Mesures oculométriques

Une autre méthode, intuitive dans le cadre de la recherche visuelle, consiste à suivre le regard des participants. En effet, des études ont montré que l'attention visuo-spatiale joue un rôle important dans la survenue de saccades oculaires. Shepherd et al. (1986) ont ainsi montré que des participants sont plus rapides pour détecter une cible si elle est située au même endroit que l'endroit ciblé par une saccade dont la direction est imposée par un signal

avant l'apparition de la cible. Hoffman and Subramaniam (1995) ont montré que les individus ne peuvent pas tourner leur regard dans une direction et concentrer leur attention dans une autre direction. En effet, le taux de détection d'une cible (une lettre dans cette étude) était dégradé lorsque l'on demandait aux participants de réaliser une saccade dans une direction autre que celle de la cible juste après la présentation de cette dernière. Il apparaît ainsi que les saccades oculaires peuvent être de bons témoins de l'orientation de l'attention visuelle dans une direction de l'espace. Theeuwes et al. (1998) ont ainsi observé que des participants qui réalisaient une saccade en direction d'une cible de couleur différente pouvaient être interrompus dans leur mouvement oculaire par un nouvel objet non pertinent. Une saccade non volontaire en direction du nouvel objet était observée. Une capture attentionnelle peut ainsi être mise en évidence par la survenue de saccades oculaires orientées vers un objet qui attire l'attention.

Le suivi du regard présente l'avantage de pouvoir être utilisé dans des environnements plus proches de la réalité, comme par exemple par Tardieu et al. (2015) pour mesurer la qualité d'un système de sonification d'une interface de navigation d'un véhicule. La présence de ce système de sonification se traduit par une diminution de la durée et du nombre de saccades vers l'interface et donc une augmentation de la durée de fixation sur la tâche principale (la conduite) par rapport à la tâche secondaire (la navigation dans l'interface).

2.1.2 Mesure de capture attentionnelle dans la modalité auditive

L'étude de la capture attentionnelle dans la modalité auditive a fait l'objet d'une revue de littérature par Dalton and Hughes (2014).

2.1.2.1 Capture attentionnelle auditive explicite

De manière similaire au phénomène de cécité attentionnelle dans la modalité visuelle, on parle de surdité attentionnelle dans la modalité auditive. Mack and Rock (1998a) ont montré en premier ce phénomène : leurs participants devaient écouter une séquence de lettres à mémoriser dans une oreille et chercher à détecter la lettre "A" dans cette séquence. Ils ne remarquaient alors pas un son diffusé dans l'autre oreille, alors qu'ils l'entendaient systématiquement une fois mis dans un état d'attention complète. Dalton and Fraenkel (2012) ont inclus dans une scène sonore un personnage répétant "I'm a gorilla" pour reprendre l'expérience de Simons and Chabris (1999) dans la modalité auditive. Les participants détectaient presque tous cet intrus dans le cas où on leur demande de prêter attention à tout ce qui pouvait paraître inhabituel (condition contrôle), alors que dans une condition où on leur demandait de prêter attention à une conversation en particulier dans la scène, une majorité ne remarquait pas l'intrus, même quand celui-ci était situé spatialement au même endroit que la conversation à suivre. Dans leurs travaux, Murphy et al. (2017) ont également mis en évidence ce phénomène de surdité attentionnelle avec un mot prononcé sur un fond de bruit blanc dans une oreille, alors que les participants recherchaient une cible (un son pur à basse fréquence) dans l'autre oreille. Koreimann et al. (2014) ont montré que des musiciens experts pouvaient ne pas remarquer un solo de guitare électrique ajouté dans un morceau de musique classique (un extrait de *Ainsi parlait Zarathustra*, de Richard Strauss) quand ils devaient réaliser une tâche qui les mobilisait sur un autre aspect de la musique (compter le nombre de coups de timbale). Enfin, Dehais et al. (2014) ont étudié ce phénomène de surdité attentionnelle sur des pilotes dans un cockpit d'avion. Les participants soumis à des conditions de vol agité (conditions météorologiques critiques) réagissaient moins ou ne remarquaient pas une alarme sonore par rapport aux conditions de vol calmes.

2.1.2.2 Capture attentionnelle auditive implicite

Comme dans la modalité visuelle, il existe dans la modalité auditive des méthodes implicites permettant d'observer le phénomène de capture attentionnelle.

Irrelevant Sound Effect

L'"irrelevant speech effect" correspond à l'observation de la dégradation de la performance d'individus lors d'une tâche de mémorisation d'items (nombres, lettres) lorsqu'un discours est diffusé en fond sonore pendant la présentation des items (Colle and Welsh, 1976 ; Ellermeier and Zimmer, 1997 ; Salame and Baddeley, 1982). L'"irrelevant speech effect" a été rebaptisé "irrelevant sound effect" après que l'on a découvert que l'effet était aussi présent en présence d'un flux audio autre qu'un discours. Différentes caractéristiques de ce flux audio ont été testées : musique (Ellermeier and Hellbrück, 1998 ; Salamé and Baddeley, 1989), variations de tonalité (Jones and Macken, 1993), variations de niveau (Tremblay and Jones, 1999).

Le paradigme de tâche de mémorisation pour mettre en évidence un "irrelevant sound effect" est intéressant mais présente des contraintes et des limites. À chaque essai, les participants doivent mémoriser 10 items (à la cadence d'un item par seconde généralement dans les études), les garder en mémoire quelques secondes puis les restituer. Ce type de passation est beaucoup plus coûteux en temps que celui des autres paradigmes, et beaucoup plus fatigant pour le participant.
De plus, ce paradigme repose sur l'observation d'une dégradation de performance au niveau de la mémorisation d'items. Or, les liens entre attention et mémoire relèvent d'interactions complexes et de plus haut-niveau que la capture attentionnelle seule. D'après Engle (2002) par exemple, plus la capacité de mémoire de travail est grande, plus la capacité à contrôler son attention est grande également. Les capacités attentionnelles seraient en effet liées aux capacités de maintien actif des représentations en mémoire de travail (Kane and Engle, 2003).

Avec ce paradigme, on n'a pas directement accès à l'effet de la capture attentionnelle, mais à celui de toute une chaîne de traitements cognitifs allant de la perception à la restitution en passant par l'encodage et le maintien en mémoire de travail (Miles et al., 1991). On peut de plus reprocher à ce paradigme de ne mettre en évidence que les stimuli sonores non pertinents qui vont jusqu'à la perturbation du maintien en mémoire de travail, mais de ne pas permettre de relever ceux qui capturent l'attention mais sans perturber suffisamment les participants pour observer une dégradation de leur performance mémorielle.
Enfin, dans ce paradigme, les items sont présentés dans la modalité visuelle, ce qui implique une interaction inter-modale et le partage des ressources attentionnelles dans ces deux modalités. Or, nous souhaitons d'abord mettre en évidence le phénomène dans la modalité auditive seule avant d'éventuellement nous intéresser aux effets inter-modaux.

Paradigme du singleton additionnel

Le paradigme du singleton additionnel a été adapté dans la modalité auditive pour la première fois par Dalton and Lavie (2004). Un effet de capture attentionnelle par un singleton y a été observé dans des séquences de 5 items sonores. En effet, les temps de réponse obtenus étaient significativement plus grands pour les séquences où un singleton était présent. Les cibles étaient définies par une durée, intensité ou fréquence particulière, et le singleton par une dimension qui différait de celle permettant de rendre la cible particulière. Par exemple, les participants devaient discriminer la cible qui avait un niveau plus ou moins élevé que les distracteurs. Dans certaines séquences, un singleton à une fréquence différente de la fréquence commune à tous les autres sons était présent . Des résultats similaires à ceux de la modalité visuelle ont ensuite été trouvés dans une seconde étude (Dalton and Lavie, 2007), dans des conditions semblables à leur première expérience, à quelques ajustements près (durée des stimuli, dimension de la cible et du singleton).

Les détails de ce paradigme seront précisés dans la suite du chapitre, lorsqu'il s'agira de l'adapter à nos travaux.

Autres paradigmes

Bidet-Caulet et al. (2015) ont proposé un nouveau paradigme destiné à sonder des processus bottom-up et top-down de l'attention auditive. Dans leur paradigme, les participants doivent détecter une cible, auditive ou visuelle. Pendant le laps de temps qui précède l'apparition de la cible, un distracteur sonore non attendu peut être diffusé. La durée qui sépare ce distracteur de la cible est variée, et le temps de capture attentionnelle est déduit des différences de temps de détection de la cible dans les différentes conditions. L'effet de processus top-down est également mesuré, car un indice informatif ou non-informatif est donné en amont. Ce paradigme a également été combiné avec des mesures d'électro-encéphalographie (EEG). Il ne permet cependant pas l'étude de la capture attentionnelle par des sons dans un flux temporel, mais seulement par des sons isolés.

Les différents paradigmes de mesure de la capture de l'attention que l'on peut mettre en oeuvre ou imaginer peuvent parfois être combinés avec des mesures physiologiques sur les participants. Il faut pour cela identifier des marqueurs qui sont corrélés à la capture de l'attention. Deux sont principalement utilisés : les mesures EEG et la pupillométrie.
Kaya et al. (2020) ont ainsi observé l'effet du timbre, de la hauteur et de l'intensité de notes de musique sur la capture attentionnelle en mesurant des réponses cérébrales pendant la diffusion des stimuli. Bidet-Caulet et al. (2015) ont relevé la modulation de certaines composantes de la réponse cérébrale (N1 et P3) par les processus top-down et bottom-up de l'attention auditive.
Enfin, des études (Boswijk et al., 2020 ; Liao et al., 2016) ont montré que la dilatation de la pupille pourrait être un autre marqueur de la saillance d'évènements sonores.

Modalité	Mesure	Méthode expérimentale
visuelle	explicite	paradigmes de cécité attentionnelle
visuelle	implicite	paradigme du singleton additionnel
visuelle	implicite	paradigme de caractéristique non pertinente
visuelle	implicite	paradigme d'amorçage
visuelle	implicite	paradigmes avec mesures oculométriques
auditive	explicite	paradigmes de surdité attentionnelle
auditive	implicite	paradigme du singleton additionnel
auditive	implicite	paradigme de l'irrelevant sound effect
auditive	implicite	paradigme proposé par Bidet-Caulet et al. (2015)
auditive	implicite	paradigmes avec mesures physiologiques (EEG, pupillométrie)

Table 2.1 – Résumé des différentes méthodes expérimentales pouvant être envisagées pour mettre en évidence un effet de capture attentionnelle en vision et en audition.

Nous récapitulons les différentes méthodes expérimentales envisageables pour mesurer un phénomène de capture attentionnelle en vision et en audition au tableau 2.1.

2.1.2.3 Discussion sur la sélection d'un paradigme

Dans ces travaux, nous souhaitons mettre en évidence un phénomène de capture attentionnelle, pour pouvoir ensuite en étudier une potentielle modulation par des propriétés sonores de notre choix. Il nous faut donc choisir une méthode adaptée à ce besoin. Nous détaillons ici les critères qui ont motivé notre choix.

Premièrement, les contraintes liées à la mise en oeuvre et l'étude du phénomène de capture attentionnelle en laboratoire orientent davantage vers une méthode implicite qu'une méthode explicite. En effet, une fois les participants conscients du fait que certains stimuli non pertinents peuvent survenir, le phénomène de cécité attentionnelle est remis en cause. Par exemple, une fois

que l'on demande aux participants s'ils n'ont rien remarqué d'étrange dans la vidéo de Simons and Chabris (1999) (cf. 1.2.4), ils deviennent beaucoup plus vigilants et ne se laissent plus surprendre. Ainsi, chaque participant ne peut participer qu'à un nombre réduit d'essais, tant qu'il ne sait pas que certains évènements inattendus peuvent survenir. Il faut donc un grand nombre de participants pour compenser ce faible nombre d'essais par personne. Par exemple, Most et al. (2000) ont fait passer 145 participants, chacun réalisant 5 essais, Newby and Rock (1998) ont recruté 96 participants, chacun réalisant 7 essais. Ce format se prête mal à l'étude de l'effet de variations de différents paramètres sonores sur la saillance, car le nombre de participants requis deviendrait très important (à chaque changement de nature ou de valeur du paramètre testé, il faudrait de nouveaux participants en nombre important).

Par ailleurs, il faut faire un choix parmi les différents paradigmes présentés ci-dessus en matière de capture attentionnelle implicite. Ceux-ci visent à observer à quel point des participants peuvent ignorer des stimuli qu'ils savent non pertinents pour la réalisation d'une tâche principale. Des adaptations des paradigmes dans la modalité visuelle présentés en 2.1.1.2 pourraient être envisagées.

Pour le paradigme de caractéristique non pertinente, ce sont principalement les stimuli qui surgissent soudainement qui capturent l'attention, et pas ceux qui se démarquent par une caractéristique statique (Hillstrom and Yantis, 1994 ; Jonides and Yantis, 1988 ; Remington et al., 1992). De plus, on note que ce dernier est coûteux à mettre en oeuvre. En effet, il nécessite de soumettre chaque participant à des essais avec plusieurs nombres d'items différents, et d'avoir assez de passages à chaque nombre d'items pour étudier les temps de réaction dans les deux conditions (cible rendue particulière ou non), pour ensuite pouvoir obtenir la pente de la droite liant temps de réponse et nombre d'items. Nous ne retiendrons donc pas le paradigme de caractéristique non pertinente pour la suite.

Le paradigme d'amorçage (Folk et al., 1992,9 ; Gibson and Kelsey, 1998) présente lui des difficultés lorsqu'il s'agit de le mettre en place dans la modalité auditive. En effet, l'amorçage a lieu un instant avant la présentation des items. Cela implique l'utilisation d'une dimension, le temps, dédiée à la mise en place de l'amorce. Dans la modalité visuelle, les items sont ensuite présentés en étant répartis dans l'espace. Ainsi l'amorçage peut bien se faire sur une dimension, le temps, et la présentation sur une autre dimension, l'espace. Dans la modalité auditive, la question de la dimension utilisée pour la présentation des items se pose (voir section 2.2.1) : généralement, la présentation des items sonores se fait dans le temps plutôt que dans l'espace. Ceci est donc incompatible avec la nécessité de présenter une amorce un instant avant les items. Nous ne retiendrons donc pas non plus le paradigme d'amorçage.

Le paradigme proposé par Bidet-Caulet et al. (2015) pourrait être intéressant à mettre en oeuvre, en se focalisant sur la partie de son design qui permet de mesurer les effets bottom-up de l'attention, mais il ne permet pas d'étudier le phénomène de capture attentionnelle par des sons situés dans des flux auditifs. Or, la saillance se définit forcément dans un contexte, donc un flux auditif donné (voir section 1.2.1). Nous ne retiendrons donc pas non plus ce paradigme.

Le paradigme du singleton additionnel ne souffre pas des différentes limites évoquées précédemment. Il semble d'une part nécessiter de la part des participants un état d'attention dans lequel les caractéristiques statiques des stimuli peuvent donner lieu à une capture attentionnelle (Pashler, 1988 ; Theeuwes, 1992,9) (contrairement au paradigme de caractéristique non pertinente), et d'autre part être compatible avec le mode de présentation des items imposé par la modalité auditive (contrairement au paradigme d'amorçage). C'est ce paradigme qui a été choisi pour les premières tentatives d'adaptation dans la modalité auditive (Dalton and Lavie, 2004,0). L'effet étant bien observé dans ces premières adaptations, c'est naturellement vers ce paradigme

que notre choix s'est orienté.

Les mesures physiologiques sont également intéressantes mais nécessitent un contrôle très précis de l'état des participants et de la diffusion des stimuli, et sont plus lourdes à mettre en oeuvre expérimentalement. Elles pourraient à long-terme être utilisées en complément sur le paradigme retenu.

Les arguments exposés dans la partie précédente ont orienté notre choix vers le paradigme du singleton additionnel, dont nous détaillons maintenant l'adaptation que nous en avons faite.

2.2 Évaluation du paradigme du singleton additionnel

Nous nous sommes d'abord posés la question du mode de présentation des items utilisés (répartition spatiale ou temporelle notamment). Nous avons ensuite reproduit l'expérience de Dalton and Lavie (2004) et analysé les résultats afin d'avoir le contrôle sur le phénomène mis en évidence.

2.2.1 Choix du mode de présentation des items

Pour pouvoir retranscrire ce paradigme dans le domaine auditif, il faut pouvoir présenter un certain nombre d'items à un participant et lui faire chercher une cible au sein de ces items. La question de l'espace dans lequel on souhaite séparer les items les uns des autres s'est alors posée. En effet, comme décrit plus haut, pour que le phénomène mis en évidence soit bien celui de la capture attentionnelle, il faut que les participants cherchent une cible dans un ensemble d'items identiques et distincts.

La question de la distinction des items pose moins de difficulté dans la modalité visuelle car, en les présentant à divers emplacements dans l'espace, il est aisé de les rendre clairement séparés les uns des autres. Dans la modalité

auditive en revanche, rien ne garantit qu'il soit possible de présenter différents stimuli et qu'ils soient bien perçus distinctement les uns des autres. Or, il faut éviter le cas de la figure 2.5 où les items se "superposent" et sont perçus comme une mixture et non comme plusieurs objets séparés.

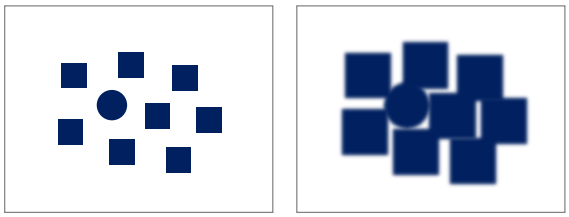

FIGURE 2.5 – Exemple de présentation visuelle d'items distincts (à gauche) et non distincts (à droite).

Nous avons donc du vérifier en amont si des participants étaient capables d'entendre distinctement plusieurs sons diffusés simultanément dans l'espace. Il fallait en effet pouvoir présenter 5 sons simultanément pour présenter une cible, un singleton additionnel, et trois distracteurs identiques (cf. section 2.1.1.2).

Les expériences ayant pour but de vérifier cette hypothèse sont présentées en annexe 1. Nous avons conclu de ces expériences que des participants ne pouvaient pas distinguer plus de 3 sons simultanés distinctement dans l'espace.

Nous avons donc choisi de distinguer les sons sur la dimension temporelle. Ce choix d'une présentation des items sonores selon cette dimension est par ailleurs conforté par les travaux de Kubovy and Van Valkenburg (2001) et Kubovy (2017) qui ont fait valoir que la fréquence et le temps étaient les deux attributs indispensables de l'audition, en analogie avec le temps et l'espace dans le domaine visuel ("Theory of Indispensable Attributes"). Justus and List (2005) ont d'ailleurs retenu ces dimensions pour faire le parallèle, dans la modalité auditive, d'un paradigme issu de la modalité visuelle visant à comparer deux modes de traitement de l'information, l'un de manière locale l'autre de manière globale. Cette question du mode de traitement de l'information a

depuis été abordée par des expériences (Susini et al., 2020) dans lesquelles les stimuli sont présentés selon les dimensions [temps, fréquence] plutôt que [temps, espace]. Cette idée d'un parallèle entre [temps, espace] en vision et [temps, fréquence] en audition est en accord avec les observations faites sur l'organisation spatiotopique de la rétine et du cortex visuel primaire, alors que la cochlée et le cortex auditif primaire sont organisés de manière tonotopique (Lauter et al., 1985 ; Merzenich et al., 1982 ; Pantev et al., 1988 ; Romani et al., 1982 ; Talavage et al., 2004).

2.2.2 Reproduction du paradigme de Dalton and Lavie (2004)

L'adaptation du paradigme du singleton additionnel dans la modalité auditive en répartissant les différents items dans le temps a été menée par Dalton and Lavie (2004). Dans leur étude, les auteures ont révélé un effet de la présence d'un singleton sur les temps de réponse pour détecter une cible parmi 5 items, mettant ainsi en évidence un phénomène de capture attentionnelle auditive par un singleton.

La première expérience que nous avons menée a consisté en la réplication de leur expérience. L'objectif était de mettre en place le protocole expérimental et de vérifier que nous obtenions des résultats similaires à ceux obtenus dans leurs travaux. Le paradigme doit en effet être réplicable si l'on souhaite observer puis faire varier un phénomène de capture attentionnelle.
La reproduction de cette expérience visait également à définir et affiner le cadre dans lequel les expériences futures seraient menées pour mettre en évidence la capture attentionnelle et son lien avec différents paramètres acoustiques ou psychoacoustiques.

La tâche à réaliser consistait à rechercher et identifier une cible sonore parmi d'autres stimuli sonores. La cible en question était définie par son niveau, différent de celui des autres sons, comme dans l'expérience notée 4A dans l'étude de Dalton and Lavie (2004). Cette cible devant se démarquer par une propriété unique, tous les autres sons étaient au même niveau. Les

participants devaient identifier si la cible était plus faible ou plus forte que les autres sons. Il s'agissait donc d'une tâche de discrimination. Ce type de tâche a été privilégié par rapport à une tache de détection car on s'assurait alors que l'attention était bien mobilisée (Dalton and Lavie, 2007). Cette tâche pouvait être réalisée en présence ou non d'un singleton. Ce singleton était un son identique aux distracteurs, sauf sur une dimension particulière (différente de la cible), la hauteur, où il prenait une valeur qui le rendait unique. Nous cherchions à voir si le facteur "présence du singleton" avait un effet significatif sur les deux variables dépendantes que sont le temps de réponse et le taux d'erreurs. Si tel était le cas, le phénomène de capture attentionnelle serait alors mis en évidence, comme ont pu le montrer Dalton et Lavie.

Participants

12 participants ont pris part à cette expérience (6 hommes, 6 femmes). Ils étaient tous consentants et âgés de 16 à 53 ans (moyenne : 29 $\pm$ 10 ans) et ont tous fait part d'une audition normale. Il n'ont pas reçu d'indemnisation.

Équipement

L'expérience a été conçue et s'est déroulée sur un MacBook pro (2020), avec le logiciel Max (version 8). Les stimuli ont été conçus avec le logiciel Ableton Live (version 10) [1]. Les stimuli étaient présentés durant l'expérience dans un casque Beyerdynamic 770 Pro (250 Ohms).

Stimuli

Les distracteurs étaient des tons purs (sinusoïdes) de fréquence 440 Hz et de niveau de pression acoustique 79 dB SPL, mesuré en sortie de casque à l'aide d'un sonomètre 2238 médiateur de la marque Brüel & Kjaer. La cible forte était un ton pur de fréquence 440 Hz et de niveau 84 dB SPL. La cible faible était un ton pur de fréquence 440 Hz et de niveau 73 dB SPL. Le

1. https ://www.ableton.com/

singleton était un ton pur de fréquence 520 Hz et de niveau 79 dB SPL. Ces valeurs ont été prises pour être au plus proche du paradigme de Dalton and Lavie (2004). Chaque item durait 100 ms et était séparé du suivant par un intervalle de 50 ms, l'IOI ("Inter-Onset Interval") était donc de 150 ms (voir figure 2.6). Une séquence de 5 items durait 700 ms au total.

Chaque participant passait 600 essais, par séries de 100. Chaque condition avec une valeur de niveau de la cible (plus fort ou plus faible) et une valeur de présence du singleton plus aigu (présent ou absent) était présentée dans 25% des cas. Le premier son de chaque séquence n'était jamais une cible ou un singleton. La cible pouvait se trouver en 3ème ou 4ème position. Quand il y avait un singleton en plus de la cible, il était situé soit juste avant, soit juste après la cible.

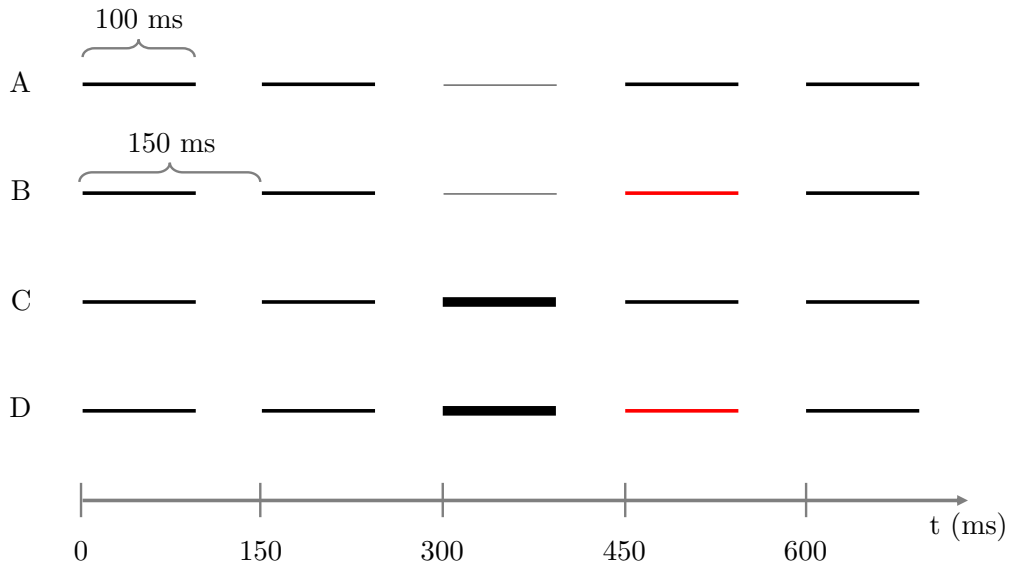

FIGURE 2.6 – Exemple de stimuli avec la cible en 3ème position. A : cible faible en absence de singleton, B : cible faible en présence de singleton après la cible, C : cible forte en absence de singleton, D : cible forte en présence de singleton après la cible. Le singleton est en rouge, la cible faible en trait fin, la cible forte en trait épais).

Procédure

Il était demandé au participant, au début de l'expérience, de se concentrer sur le niveau sonore des sons afin de détecter la cible et de juger si celle-ci était plus faible ou plus forte que les autres sons. Il était précisé de ne pas prêter attention aux sons plus aigus qui pourraient survenir. Le participant devait donc mener une discrimination entre deux cibles, d'un niveau différent de celui des autres sons.

Un essai se déroulait de la manière suivante : l'inscription « prêt ? » s'affichait à l'écran durant 500 millisecondes, puis l'inscription « Go ! » s'affichait et une séquence était diffusée. Le protocole de réponse était un protocole de type choix forcé à 2 alternatives. Ainsi, à la fin de la séquence, les boutons « faible » et « fort » s'affichaient et le participant pouvait répondre. Le message « correct » ou « incorrect » était donné en fonction de l'exactitude de la réponse si une réponse avait été donnée. Si après 3000 ms aucune réponse n'était donnée par le participant, le message « Trop tard. Répondez plus vite ! » était affiché. Ce message était affiché 1500 ms, puis le message « prêt ? » qui annonçait l'essai suivant était affiché. Le temps de réponse était mesuré entre la fin de la séquence et l'instant où une réponse était donnée.

Chaque participant débutait l'expérience par une phase d'entraînement. L'expérimentateur était présent durant l'entraînement pour donner d'éventuelles explications, répondre à des questions et s'assurer que tout se déroule correctement. Lorsque le participant parvenait à enchaîner les essais et obtenaient strictement plus de 60% de réponses correctes, l'entraînement était validé. L'expérience durait 1 heure.

Résultats

Pour l'analyse des taux d'erreurs, les essais pour lesquels une réponse a bien été donnée ont été retenus, soit 99,7% des données. Pour l'analyse des temps de réponse, les données où la réponse était correcte ont été gardées,

soit 81,8% des données.
Pour les deux variables dépendantes (temps de réponse et taux d'erreurs), le facteur présence du singleton (présent ou absent) a d'abord été examiné. Les résultats son présentés au tableau 2.2. L'effet de la présence du singleton est d'augmenter les temps de réponse et les taux d'erreurs. Des test-t appariés (pour prendre en compte le fait que les mêmes participants répondent en absence ou présence de singleton) ont révélé un effet significatif du facteur "présence du singleton" sur les temps de réponse (T(11) = -4,260, p < 0,001, cohen-d = 0,75, power = 0,78) et les taux d'erreur (T(11) = -7,732, p < 0,001, cohen-d = 1,23, power = 0,99).

	Singleton absent	**Singleton présent**
Temps de réponse	521 ms	667 ms
Taux d'erreurs	13,8 %	22,8%

TABLE 2.2 – Temps de réponse et taux d'erreurs moyens en fonction de la présence du singleton

Les séquences où un singleton est présent ont ensuite été conservées, et pour les temps de réponse comme pour les taux d'erreurs, une ANOVA à mesures répétées a été menée sur ces séquences, pour les deux facteurs : Niveau (2 niveaux de cible : faible ou fort) $\times$ Position (2 positions de singleton : avant ou après la cible)(voir tableaux 2.3 et 2.4).
Cette analyse a révélé un effet significatif de la position du singleton pour les temps de réponse (MSE = 231.10^3, F(1,11) = 11,3, p < 0,01, η^2 = 0,09) [2] et pour les taux d'erreurs (MSE = $2{,}30.10^3$, F(1,11) = 83,7, p < .001, η^2 = 0,19). L'analyse a également révélé un effet significatif du niveau de la cible pour les taux d'erreurs (MSE = $2{,}34.10^3$, F(1,11) = 17,9, p < .01, η^2 = 0,20). Enfin, un effet significatif de l'interaction Position du singleton $\times$ Niveau de la cible pour les temps de réponse (MSE = $74{,}9.10^3$, F(1,11) = 14,1, p < .01, η^2 = 0,029) et pour les taux d'erreurs (MSE = $2{,}30.10^3$, F(1,11) = 60,0, p < .001, η^2 = 0,19) a été révélé.

2. MSE :"Mean Square Error", F : statistique F, p : "p-value", η^2 : taille de l'effet

Source	SS	ddl	MSE	F	p	η^2
Niveau	$59{,}1.10^3$	$(1, 11)$	$59{,}1.10^3$	4,13	0,067	0,023
Position	231.10^3	$(1, 11)$	231.10^3	11,3	0,006	0,090
Niveau * Position	$74{,}9.10^3$	$(1, 11)$	$74{,}9.10^3$	14,1	0,003	0,029

TABLE 2.3 – ANOVA avec les facteurs niveau de la cible (niveau) et position du singleton (position) pour les temps de réponse. SS : "Sum of Squares", ddl : degrés de liberté, MSE : "Mean Square Error", F : statistique F, p : "p-value", η^2 : taille de l'effet.

Source	SS	ddl	MSE	F	p	η^2
Niveau	$2{,}34.10^3$	$(1, 11)$	$2{,}34.10^3$	17,9	0,001	0,20
Position	$2{,}30.10^3$	$(1, 11)$	$2{,}30.10^3$	83,7	$< 0{,}001$	0,19
Niveau * Position	$2{,}30.10^3$	$(1, 11)$	$2{,}30.10^3$	60,0	$< 0{,}001$	0,19

TABLE 2.4 – ANOVA avec les facteurs niveau de la cible (niveau) et position du singleton (position) pour les taux d'erreurs. SS : "Sum of Squares", ddl : degrés de liberté, MSE : "Mean Square Error", F : statistique F, p : "p-value", η^2 : taille de l'effet.

L'interaction significative entre les 2 facteurs indique que l'effet du singleton est différent selon sa position, et l'effet de sa position dépend du niveau de la cible. Une analyse post-hoc (test HSD de Tukey) a révélé que l'effet de cette interaction était essentiellement présent quand la cible était au niveau faible et que le singleton survenait juste avant celle-ci (voir figures 2.7 et 2.8).

Discussion

Les résultats obtenus semblent bien témoigner d'un phénomène de capture attentionnelle. Le temps de réponse moyen est plus grand en présence du singleton qu'en son absence, et l'intensité de l'effet de sa présence est fort (cohen-d = 0,75). Les temps de réponse moyens obtenus (521 ms en absence de singleton, 667 ms en présence de celui-ci) sont plus longs que ceux obtenus par Dalton et Lavie (257 ms en absence de singleton, 312 ms en présence de celui-ci). La présence du singleton se traduit par un allongement du temps de réponse de 21,4% dans leur étude, de 28,0% dans notre réplication.

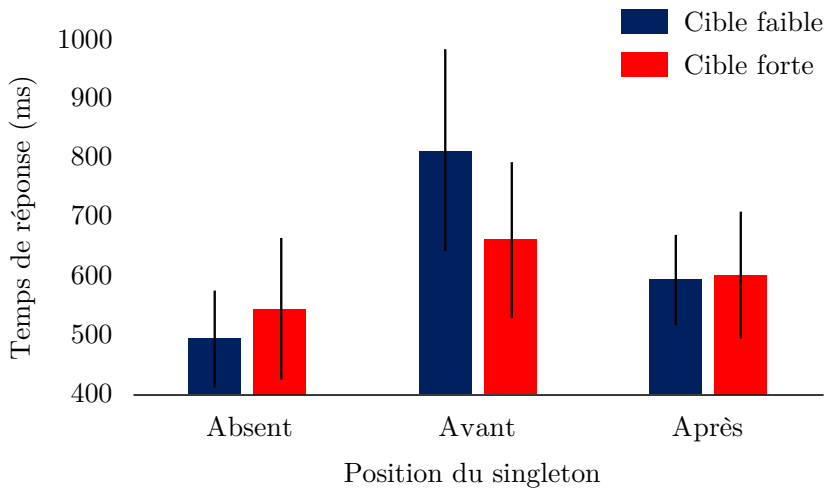

FIGURE 2.7 – Temps de réponse (ms) pour l'interaction (niveau de la cible) × (position du singleton) (cible faible en bleu, cible forte en rouge, singleton absent à gauche, singleton avant la cible au milieu, singleton après la cible à droite). Les barres verticales représentent les intervalles de confiance à 95%.

La présence du singleton a également un fort effet sur les taux d'erreurs (cohen-d = 1,23). Comme pour les temps de réponse, les taux d'erreurs sont un peu plus élevés dans notre expérience que dans celle de Dalton and Lavie (2004) : 10% en absence du singleton et 17% en sa présence dans leur expérience, contre 13,8% et 22,8% dans notre réplication. L'augmentation du taux d'erreurs relative est de 70% dans leur cas, et de 65% dans notre réplication. Les résultats confirment dans tous les cas que la tâche est assez facilement réalisable : on observe en effet des taux d'erreurs de 13,8% en absence de singleton.

La différence significative des temps de réponse et des taux d'erreurs observée entre la condition où la cible faible survient après le singleton et les autres conditions suggère la présence d'un phénomène de masquage temporel au sein des séquences. La cible faible serait masquée par le singleton, plus que par les autres sons. La différence de sonie entre le distracteur et le singleton expliquerait cette différence entre la condition cible faible après un distracteur et cible faible après le singleton. De fait, pour des fréquences de l'ordre de celles utilisées dans l'expérience (440-520 Hz), les sons plus aigus sont perçus

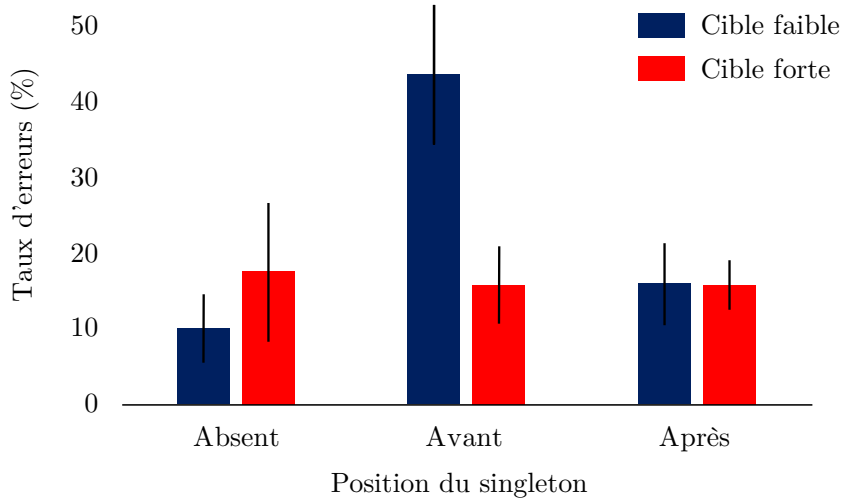

FIGURE 2.8 – Taux d'erreur (%) pour l'interaction (niveau de la cible) × (position du singleton) (cible faible en bleu, cible forte en rouge, singleton absent à gauche, singleton avant la cible au milieu, singleton après la cible à droite). Les barres verticales représentent les intervalles de confiance à 95%.

comme plus forts s'ils ne sont pas égalisés en sonie (Fletcher and Munson, 1933). Les sonies spécifiques de chaque item ont donc été calculées (voir figure 2.9). Les sonies globales associées (méthode de Zwicker) sont de 19,3 sones pour la cible faible, 28,6 sones pour le distracteur et de 31,3 sones pour le singleton. Le singleton était donc perçu plus fort que le distracteur, et pouvait donc potentiellement donner lieu à un phénomène de masquage.

Il convient donc de prendre en compte cet aspect dans la suite des expériences, où les effets de masquage ne doivent pas interférer avec l'effet de capture attentionnelle.
Cependant, les temps de réponse sont supérieurs à 590 ms pour toutes les séquences contenant un singleton, même celles où le phénomène de masquage n'a pas lieu. Le temps de réponse moyen pour les séquences où le singleton était absent est de 521 ms. Ainsi, si l'effet est significativement plus marqué pour une cible faible qui survenait après le singleton, il a été noté que ces temps de réponse restaient toujours significativement plus longs pour toutes les séquences où un singleton était présent que pour les séquences où

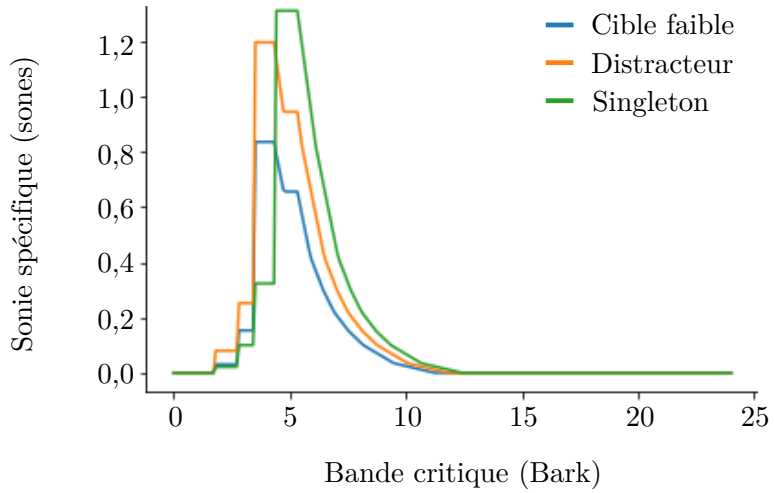

FIGURE 2.9 – Sonie spécifique de la cible faible (en bleu), du distracteur (en jaune) et du singleton (en rouge) en fonction des bandes critiques.

le singleton était absent. Un effet de la présence du singleton a donc bien été observé grâce à cette expérience, en plus du potentiel effet de masquage.

Le paradigme semble ainsi adapté pour mettre en évidence un phénomène de capture attentionnelle par un singleton. Nous avons souhaité l'ajuster pour dépasser les limites soulevées ci-dessus (effet de masquage notamment) et pouvoir l'utiliser pour tester l'effet de différents propriétés acoustiques du singleton sur la capture attentionnelle.

2.2.3 Adaptation du paradigme du singleton additionnel

Égalisation en sonie

La reproduction de l'expérience de Dalton and Lavie (2004) l'a montré : il faut être vigilant lors de la présentation d'une succession de sons car un effet de masquage peut se produire. Dès lors, nous avons choisi, pour la suite des expériences, d'égaliser en sonie tous les sons utilisés dans les séquences présentées aux participants. Cette contrainte nous était de toute manière imposée par notre volonté d'étudier l'influence des paramètres acoustiques ou psychoacoustiques, notamment les paramètres du timbre, influant la saillance

d'un son à sonie égale.

Variations de la fréquence fondamentale des items

L'expérience menée précédemment contenait plusieurs séries, chacune contenant plusieurs dizaines d'essais, chaque essai consistant en l'écoute d'une séquence de 5 sons. Cette expérience était donc répétitive, les séquences étant assez semblables les unes aux autres (les mêmes distracteurs sont utilisés d'un essai à l'autre). Seules la position de la cible, la présence d'un singleton et sa position quand il était présent (avant ou après la cible) pouvaient varier entre les séquences. On dénombrait ainsi seulement 12 séquences différentes, croisant les valeurs de position de la cible (2 valeurs possibles : 3 ou 4), type de cible (2 valeurs : A ou B), présence du singleton (présent ou absent), et si le singleton était présent, position du singleton (2 valeurs possibles : avant ou après la cible) (voir figure 2.10).

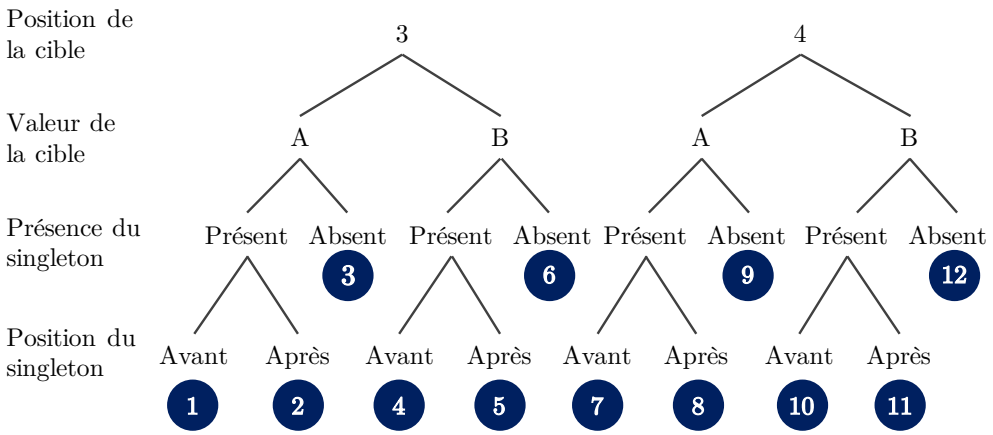

FIGURE 2.10 – Dénombrement des stimuli de l'expérience pilote avec un singleton. Le croisement des facteurs position de la cible (3 ou 4), valeur de la cible (A ou B), présence du singleton (présent ou absent), et si le singleton est présent, position du singleton (avant ou après la cible) donne 12 stimuli différents.

Nous avons souhaité, pour la suite des expériences, éviter tout effet de mémorisation et de reconnaissance des séquences, qui peut se produire si chacune des séquences revient toujours sous le même format. En effet, dans

ce cas, les participants pourraient apprendre et reconnaître les séquences, et parvenir à répondre sans avoir à mener de recherche de la cible à proprement parler au sein de la séquence. Pour apporter des variations dans la succession des essais, la fréquence fondamentale des items des expériences suivantes a donc été définie aléatoirement sur une plage de quelques Hertz (détaillée au cas par cas dans chaque expérience). Cette plage de variations a été choisie suffisamment petite pour que la sonie de chaque stimulus n'en soit pas affectée, mais suffisamment grande pour que l'on ne perçoive pas tous les items exactement à la même hauteur. Le minimum de variation perceptible pour des sons complexes de fréquence fondamentale inférieure à 500 Hz est de 3 Hz pour les sinusoïdes pures et de 1 Hz pour les sons complexes (Kollmeier et al., 2008). Dans la première expérience par exemple, les fréquences fondamentales des items ont été tirées au hasard selon une loi de probabilité uniforme sur une plage de valeur d'une largeur de 20 Hz. Les sons se succèdent ainsi en étant perçus à des hauteurs aléatoires et très légèrement différentes. Ainsi, il n'y a jamais deux séquences identiques.

Manipulation des paramètres acoustiques des sons

Il est d'usage de distinguer 3 dimensions élémentaires lorsqu'il s'agit de caractériser un son : hauteur, durée, intensité. Les autres dimensions sont regroupées dans le concept de timbre.

Il faut considérer soigneusement les dimensions sur lesquelles des manipulations de caractéristiques des stimuli pourraient être réalisées dans notre expérience. De fait, dans ce paradigme, les distracteurs doivent être tous identiques, la cible doit être rendue particulière sur une dimension, et le singleton doit également être rendu particulier, mais sur une autre dimension. En effet, pour réaliser la tâche, les participants dirigent leur attention spécifiquement sur la dimension définissant la cible. Si cette dimension interagit avec celle définissant le singleton, cela pourrait introduire des biais.

Par exemple, la hauteur, ou fréquence fondamentale, est un paramètre qui peut permettre de rendre le singleton unique (qui est alors le seul son plus ou moins aigu). Dans ce cas, il faut être prudent et choisir une dimension qui n'interagit pas avec la hauteur pour démarquer la cible. Le choix d'une variation de timbre, qui touche aussi aux fréquences contenues dans le son, paraît plus risqué dans ce cas qu'une variation de la durée. La question de l'interaction entre hauteur et timbre a en effet été étudiée dans le domaine musical (Allen and Oxenham, 2014 ; Marozeau et al., 2003 ; Pitt, 1994) et les deux dimensions semblent pouvoir interagir. Pitt (1994) ont par exemple montré que les variations du timbre affectaient le jugement de la hauteur du son chez des participants. Dans une de ses expériences, les participants devaient classer des stimuli sur une dimension (la hauteur) pendant que l'autre dimension, non pertinente (le timbre), était manipulée. Les temps de réaction étaient allongés et les taux d'erreurs augmentés quand il fallait classer les stimuli selon leur hauteur et que leur timbre variait de manière indépendante avec la tâche principale.

Or, notre objectif est de tester les paramètres du timbre qui peuvent rendre un son saillant : en particulier, rugosité et brillance (cf. section 1.2.2). C'est donc ce paramètre qui doit rendre le singleton unique. Ayant choisi de travailler à sonie égale, la dimension intensité est imposée. Le timbre interagissant avec la hauteur, il ne reste donc qu'une dimension possible pour définir la cible : la durée. Il y aura donc une cible courte et une cible longue.

Le paradigme ayant été défini et adapté à nos besoins, nous avons pu le mettre en oeuvre. Cette mise en oeuvre et les résultats qui en ont découlé ont fait l'objet d'une publication (Bouvier et al., 2023b) que nous proposons telle quelle au lecteur en section 2.3. Dans ces travaux, une première expérience a permis de révéler qu'un singleton brillant donnait lieu à un effet de capture attentionnelle. Puis, nous avons enrichi l'expérience en variant la brillance du singleton pour observer la modulation de la capture attentionnelle par cet attribut. Le même principe a ensuite été mené sur une autre dimension du timbre : la rugosité. Enfin, nous avons croisé ces deux dimensions dans

une dernière expérience. Les résultats ont permis de montrer qu'il existe un lien entre les variations d'attributs du timbre sonore du stimulus à ignorer (le singleton) et l'effet de capture attentionnelle. Autrement dit, nous avons montré comment certaines dimensions du timbre sonore modulent la saillance auditive.

2.3 Modulation de la capture attentionnelle par des attributs du timbre

Revealing the stimulus-driven component of attention through modulations of auditory salience by timbre attributes

Baptiste Bouvier, Patrick Susini, Catherine Marquis-Favre & Nicolas Misdariis

Attention allows the listener to select relevant information from their environment, and disregard what is irrelevant. However, irrelevant stimuli sometimes manage to capture it and stand out from a scene because of bottom-up processes driven by salient stimuli. This attentional capture effect was observed using an implicit approach based on the additional singleton paradigm. In the auditory domain, it was shown that sound attributes such as intensity and frequency tend to capture attention during auditory search (cost to performance) for targets defined on a different dimension such as duration. In the present study, the authors examined whether a similar phenomenon occurs for attributes of timbre such as brightness (related to the spectral centroid) and roughness (related the amplitude modulation depth). More specifically, we revealed the relationship between the variations of these attributes and the magnitude of the attentional capture effect. In experiment 1, the occurrence of a brighter sound (higher spectral centroid) embedded in sequences of successive tones produced significant search costs. In experiments 2 and 3, different values of brightness and roughness confirmed that attention capture is monotonically driven by the sound features. In experiment 4, the effect was found to be symmetrical : positive or negative, the same difference in brightness had the same negative effect on performance. Experiment 5 suggested that the effect produced by the variations of the two attributes is additive. This work provides a methodology for quantifying the bottom-up component of attention and brings new insights on attention capture and auditory salience.

The acoustic environment is so rich in information that our brain cannot process in detail all of the sounds it is constantly receiving. Instead, the individual selects stimuli that they deem to be relevant for a particular task, and ignores others (Desimone and Duncan, 1995). The most famous example of selective attention is the cocktail party problem (Cherry, 1953). This ability is made possible by an attentional process that filters the flow of stimulus information through certain irrelevant channels (Broadbent, 1958 ; McDermott, 2009). The precise mechanisms involved in this filtering are still being investigated (Marinato and Baldauf, 2019). However, the brain should not be completely blind to task-irrelevant stimuli since they could provide important information about the environment. For example, if we are chatting to someone on the street, we can pick up what they are saying and ignore the surrounding traffic noise. However, the squeal of tires associated with a car's sudden braking may still attract our attention. So, if the stimulus is sufficiently salient, the brain may have to process the information it contains involuntarily. This phenomenon is known as involuntary attentional capture. Salience is the property of a stimulus that makes it likely to capture attention, i.e., the bottom-up component of attention (Treue, 2003).

Attention capture has been extensively studied in the visual modality. Implicit approaches measure the behavioral costs (increased reaction times and error rates) of the presence of an irrelevant distractor in focal tasks. Among other things, irrelevant stimuli defined by their color, shape or onset time are known to attract the attention of participants performing a visual search task (Theeuwes, 1992,9 ; Yantis and Jonides, 1984).

However, there has been some debate about how salient objects can automatically capture attention. Some have argued that salient objects have an automatic power to attract attention, regardless of the subject's goals. They observed that certain features, such as color or shape, make the salient object automatically capable of attracting attention (Theeuwes, 2010). This led to a stimulus-driven conception of attentional capture (Theeuwes, 1993) : visual selection is determined by the physical properties of the stimuli, and attention is drawn to the location where one object differs from the others along a particular dimension. However, others have argued that only items that match the target's features can capture attention. For them, capture depends on the attentional set that is encouraged by the task (Folk et al., 1992). For example, it has recently been found that salience does not influence the capture of visual stimuli. Instead, participants can often learn to suppress salient objects (Gaspelin and Luck,

2018 ; Stilwell and Gaspelin, 2021). Authors from different parties eventually came together to review and compare their theories (Luck et al., 2021). They agreed that "physically salient stimuli automatically generate a priority signal that, in the absence of specific attentional control settings, will automatically capture attention, but there are circumstances under which the actual capture of attention can be prevented", reconciling the stimulus-driven and contingent capture approaches.

In the auditory modality, few studies have addressed this issue. Huang et al. (2017) used an explicit approach to measure auditory salience in complex sound scenes. Participants listened to the scenes dichotically (a different scene in each ear), and continuously indicated which side their attention was focused on. Averaged across scenes and participants, this allows the identification of salient events in a scene where their responses, on average, indicate how they orient their attention. This protocol involves top-down processes, as participants actively listen to the sounds and report the orientation of their attention. We therefore cannot infer any measurement of the purely bottom-up component of attention. In Kaya et al. (Kaya et al., 2020) the authors asked their participants to focus on a visual task and to ignore background acoustic melodies. Brain responses were recorded, showing that variations in acoustic attributes could make notes in these melodies more salient, and how these different attributes interacted to modulate brain responses.

Dalton and Lavie (2004) used an implicit approach based on the additional singleton paradigm to reveal an auditory attentional capture effect by sound features such as frequency or intensity. This paradigm was first developed in the visual modality to show that irrelevant stimuli can capture participants' attention during a visual search task, resulting in increased error rates and response times (Pashler, 1988 ; Theeuwes, 1992).

Results from Dalton and Lavie (2004) showed a significant cost (increased response times and error rates) in an auditory search task caused by irrelevant sounds. In their experiment, participants had to listen to sequences of five sounds. Among these, they had to detect a target defined by a dimension (e.g., a change in frequency compared to non-targets). In half of the trials, one of the non-targets was made different from the others on a dimension other than that which defined the target, such as intensity. This sound is called a singleton and is irrelevant to the task. In fact, paying attention to the dimension that defines the singleton is not an advantageous strategy for detecting the target. The results showed that the singleton features could cause interference : participants made more errors and took more time to

detect the target when the singleton was present. The effect was not due to low-level interactions between the singleton and the target, which would have caused it to be more difficult to compare the target with the singleton than with a non-target. The effect was shown when the singleton was separated from the target by another sound. Garrido et al. (2009) discussed the similarity to mismatch negativity studies, which focus on the elicitation of an event-related potential by deviant tones that differ in frequency or duration. The much shorter inter-stimulus interval, the frequency of occurrence of the deviant tones, and the explicit instruction to ignore these irrelevant singletons limit the parallels that can be drawn in this area of research.Dalton and Lavie (2004) focused on the attentional capture produced by singletons of different frequency or intensity, but did not investigate the effects of sounds whose features are gradually modified.

In addition, the study of variations in intensity, and therefore loudness, of sounds may be compromised in this paradigm. Masking effects are likely to occur for louder sounds and interfere with the attentional processes we wish to study (Moore, 2012). However, the paradigm is compatible with the study of variations in timbre. One precaution would be to equalize all sounds in loudness to remove potential masking effects and the influence of loudness, which can be affected by pitch or timbre variations (Melara and Marks, 1990).

None of the approaches mentioned here focused on the relationship that may exist between variations in the acoustic attributes and the attentional capture effect.

The first acoustic feature one might think of when studying salience is loudness. Sounds that are perceived as louder are more likely to attract the listener's attention. Loudness has been shown to be an important feature of salience (Huang et al., 2017 ; Kim et al., 2014 ; Liao et al., 2016). In addition to this feature, several studies have shown that some dimensions of timbre can be sound markers for conveying relevant information. Lemaitre et al. (2007) found that listeners used common perceptual dimensions to categorize car horns. Two of the three dimensions identified were roughness and brightness. Arnal et al. (2015) noted that amplitude modulated sounds in the roughness range are found in both natural and artificial alarm signals, and are better detected due to the privileged space they occupy in the communication landscape. Rough sounds are also said to enhance aversiveness through specific neural processing (Arnal et al., 2019). Brightness has long been known to be a major dimension of musical timbre (McAdams, 2019) and has therefore been included

in most salience models (Huang et al., 2017 ; Tordini et al., 2013). More recently, roughness has also been included (Kothinti et al., 2021).

Thus, the existence of the stimulus-driven component of attention capture has been theoretically established. Moreover, the additional singleton paradigm allows the measurement of the attentional capture effect due to sound features. Finally, the literature findings suggest that certain attributes of the sound timbre are potential candidates that could be responsible for the salience of a sound, and thus its ability to capture attention. However, to the authors' knowledge, no study has ever established the relationship that might exist between variations in these features and the magnitude of the attentional capture effect. In other words, the driving properties of attentional capture by the stimulus features have not yet been revealed.

In the present work, we adopted the additional singleton paradigm to provide evidence for the effect of timbre features on attentional capture. We then used this paradigm to quantify the relationship that may exist between a sound feature and the associated capture effect. Thus, in the current study, we focused on the properties of the stimulus-driving of the attentional capture effect.

To summarize, we wanted to answer the two following questions :

- Do timbre attributes such as brightness or roughness trigger attention capture ?
- How do their variations drive attention capture ?

First, the possibility of an attentional capture by a timbre variation was investigated. Therefore, the spectral centroid (SC) of the singleton, which correlates with its perceived brightness, was investigated in experiment 1. Then, the same experimental procedure was used to evaluate how the effect size was modulated by feature variations. In experiments 2 and 3, the SC and the depth of amplitude modulation (correlated with roughness) could take several different values. Finally, experiment 4 examined the effect of symmetric variations in brightness and experiment 5 focused on combined variations in brightness and roughness to investigate the directionality and additivity of attentional modulation.

Experiment 1 : attentional capture by a bright singleton

Method. *Transparency and openness.* We report how we determined our sample size, all data exclusions, all manipulations and all measures in the study. Data were collected in 2021 and 2022 and analyzed using python 3.7. All statisti-

cal analyses were performed using python 3.7 and the open-source pingouin package.

Participants. A previous pilot experiment involving 11 participants was conducted to calculate the power of the effect of the singleton presence on response time. The calculus was made for a one-tailed t-test, with an effect size of d = 0.8, $\alpha = 0.05$ and aiming for a power of 0.8, and determined a minimum sample size N = 12.

Thus, 15 participants (8 females, 7 males) took part in this experiment. They ranged in age from 20 to 45 years (mean age : 31 ± 8 years). They were all consenting and reported normal hearing. An audiometry in the frequency range between 0.125 and 8 kHz was performed for each participant and revealed no hearing impairment. The protocol was approved according to Helsinki Declaration by the Ethics Committee of Institut Européen d'Administration des Affaires (INSEAD). All methods were carried out in accordance with their guidelines and regulations. Participants gave written informed consent and received financial compensation for their participation.

Apparatus. The experiment was designed and run on Max software (version 7, https://cycling74.com), on a Mac mini 2014 (OS Big Sur 11.2.3). The stimuli were designed with python 3.7, and presented during the experiment through headphones (Beyerdynamic 770 pro, 250 Ohm). The experiment took place in the STMS laboratory of IRCAM in a soundproofed double-walled IAC booth.

Stimuli. The stimuli were made of sequences of 5 sounds (see Fig. 2.11). All notes follow the harmonic structure of Bouvier et al. (2023a), with 20 harmonics, the n^{th} harmonic f_n having a frequency $n * f_0$ and a weight $\frac{1}{n^\alpha}$. Thus, decreasing α increased the sound spectral centroid (SC), and therefore its perceived brightness : $SC = \frac{\sum_{i=1}^{20} \frac{f_i}{i^\alpha}}{\sum_{i=1}^{20} \frac{1}{i^\alpha}}$.

Distractor. For the reference distractor, $\alpha = 3$. It lasted 170 ms, with a ramp at the beginning and end of 5 ms, and had a SC equal to 512 Hz.

Targets. The targets were 50 ms shorter or longer than the distractor. This value is higher than what Abel (2015) found as a just-noticeable difference (jnd) for duration discrimination of sinusoidal sounds. Based on previous tests done in the lab, the experimenters still ensured beforehand that the targets were clearly heard as distinct from the distractors. The targets had the same fundamental frequency and spectrum

distribution ($\alpha = 3$) as the reference distractor, but a duration of 220 ms for the long one and 120 ms for the short one.

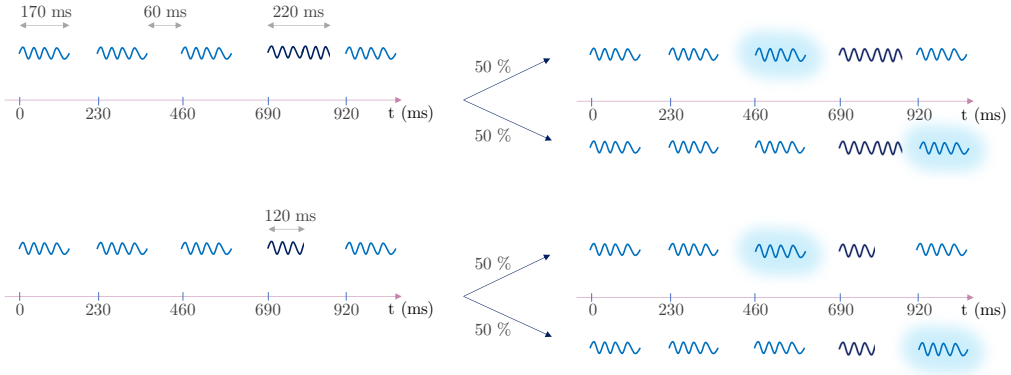

FIGURE 2.11 – Stimuli without (left) and with (right) a singleton (surrounded with a glow), with 50% chances being before or after the target (dark blue). Only sequences with target in position 4 are shown here.

Singleton. The singleton had the same fundamental frequency and envelope as the reference distractor, but a different spectrum distribution with $\alpha = 2$. It resulted in a higher SC, equal to 822 Hz. Allen and Oxenham (2014) found a jnd of 5.0% for the SC, which ensures the singleton was indeed perceived brighter. The experimenters still ensured beforehand that the singleton was clearly heard as distinct from the distractors. In the reference condition, the target was embedded in sequences of distractors only such that a sequence was composed of four distractors and a target stimulus. In the test condition, one of the distractors was the singleton such that a sequence was composed of three distractors, one target and one singleton. The IOI ("Inter-Onset Interval") was kept constant at 230 ms. The first sound of each sequence was always a distractor. The target was in 3rd or 4th position (50% of the trials each). In the trials containing a singleton, its position was either just before or just after the target (50% of the trials each). All the conditions are presented in Fig. 2.11.

Loudness equalization. All the sounds were equalized in an adjustment experiment with 12 participants from the lab, using the same setup as the main experiment. Loudness adjustments were performed by comparing all the sounds (short target, long target or singleton) to a reference (the distractor presented at 80 dB SPL).

The sounds were randomly distributed and presented 8 times each. The levels were measured at the headphones output with a Brüel and Kjaer 2238 mediator sound level meter. The obtained levels were 81 dB SPL for the short target, 79 dB SPL for the long target and 74 dB SPL for the singleton. All inter-participants standard deviations of these obtained levels were less than 1 dB SPL, i.e., less than a just-noticeable difference in sound level.

Procedure. Six blocks of 60 randomly distributed trials were run for each participant. For every trial, the word *Ready* was displayed on the screen for 1500 ms, then a sequence of 5 sounds was presented. At the end of the sequence, the participant could respond by pressing a keyboard : "1" for "short" and "2" for long (2 alternative forced choice protocol). Feedback regarding the participant's response (*Correct* or *Incorrect*) was displayed after each trial and remained for 1500 ms. If after 3000 ms no answer was given by the participant, the message *Too late. Answer faster !* was displayed. The response time was measured from the moment the target was played in the sequence. Then, a 1500 ms pause occurred and the next trial began. The participants were asked, at the beginning of the experiment, to focus on the duration of the sounds and their duration only in order to discriminate the target. Each participant had a training block before taking the test. We kept only the results of participants with an error rate below 40% on the sequences containing the target. Due to this criterion, one participant had to be replaced at this step. The experiment lasted 90 min on average.

Results. For each participant, and for each singleton condition (absent or present), we calculated the mean and the standard deviation of the response times. We then removed the data whose response time was more than two standard deviations from the mean (Miller, 1991). We also removed the data for which the response time was less than 100 ms, and those for which the participant did not answer. 94.9% of the data were kept at this stage. For the response time analysis, only the data where the participant's response was correct were kept, i.e., 75.6% of the data. The results of mean error rates and response times are presented in Table 2.5. For all the following experiments, error rates follow the same trends as response time increases. The LISAS (Linear Integrated Speed Accuracy Score - Vandierendonck (2017)) were also computed and followed the same trends. For the sake of clarity, we therefore

show only the increases in response time.

Singleton	Absent	Present
Response time	985 ms	1121 ms
(Standard deviation)	(142)	(185)
Error rate	16.2%	24.2%
(Standard deviation)	13.5	15.1

TABLE 2.5 – Mean and standard deviation of response times and error rates (across the 15 participants) depending on the presence of the bright singleton.

The error rates (16.2% and 24.2% in the conditions without and with a singleton, respectively) confirm that participants were able to complete the task correctly in both conditions. The mean response time increase, when the singleton was present, was 137 ms. A t-test revealed that the singleton presence had a significant effect on response time increase (t-test : $t(14) = 8.33$, $p < 0.001$). The effect of the singleton presence was very large (cohen-d = 2.1). A very large effect of the singleton presence was found for error rates as well ($t(14) = 3.85$, $p < 0.001$, cohen-d = 1.0).

The effect of the singleton position on error rates was not significant ($t(14) = 0.72$, $p = 0.48$), suggesting that attentional capture occurs as much whether the singleton appears before or after the target. However, there was an effect of the singleton position on response times ($t(14) = 4.38$, $p < 0.001$) : when the singleton appeared after the target, the response times were greater. This absence of effect of the singleton position on error rates and the increased reaction times when the singleton occurs after compared to before the target confirm that this effect is not due to auditory masking. This is consistent with the loudness equalization that had been carried out beforehand and the IOI which prevented auditory masking (Moore, 2012). The observed effect is due to an attentional capture caused by the bright singleton. Finally, one could claim that the effect is due to the surprise caused by the occurrence of the singleton. However, this singleton is present in 50% of the trials, and the participants identified and accustomed themselves to it during the training session. Moreover, no significant difference was found for response times between trials where a singleton appears after one or more trials without any singleton (the "surprising" condition), and trials where the singleton is present after one or more trials with a singleton (the "non-surprising" condition) : $t(14) = 0.31$, $p = 0.76$.

This first experiment thus allowed us to validate the framework in which we can test modulations of timbre features and observe how they drive the attentional capture effect. It was therefore decided to reproduce the experiment, modifying it so that the singleton could take different values of brightness in a second experiment, and different values of roughness in a third one.

Experiments 2 and 3 : variations of brightness and roughness

Experiments 2 and 3 were conducted to study how the effect magnitude is modulated by the singleton feature variations. In experiment 2, we replicated experiment 1 with four different values of the spectral centroid (SC) for the singleton. In experiment 3, four values of the amplitude modulation depth for the singleton were used. This latter sound feature is associated to an auditory attribute usually described by the semantic attribute "roughness" (Zwicker and Fastl, 2013).

Method. *Participants.* Twenty participants (10 females, 10 males) took part in experiment 2, and 20 others (10 females, 10 males) in experiment 3. The sample size was increased to ensure that the power of the effect produced by the second-brightest singleton was greater than 0.8. This was done in order to have at least two different brightness conditions with sufficient power. The participants ranged in age from 19 to 34 years (mean age : 27 ± 4 years) for experiment 2, and from 22 to 50 years (mean age : 28 ± 8 years) for experiment 3. They were all consenting and reported normal hearing. An audiometry in the frequency range between 0.125 and 8 kHz was performed for each participant and revealed no hearing impairment. Participants gave written informed consent and received financial compensation for their participation.

Apparatus. The apparatus was the same as in the first experiment, except that it took place in the INSEAD-Sorbonne Université Behavioural Lab, in soundproofed rooms.

Stimuli. The distractors and targets were the same as in experiment 1. For experiment 2, the singleton SC could take 4 values : 538, 563, 640 or 768 Hz. Each one was presented in 20% of the trials. To establish these values, an increment of SC was calculated (using the estimation of 5% for SC jnd found by Allen and Oxenham (2014), and then multiplied by 1, 2, 5 and 10. For experiment 3, the

singleton signal $s_{sing}(t)$ was the distractor signal $s_{dis}(t)$ modulated at a modulation frequency $f_{mod} = 50$ Hz : $s_{sing}(t) = (1 + m * cos(2\pi f_{mod} t)) * s_{dis}(t)$. The modulation depth m could take 4 values : 0.1, 0.2, 0.5 or 1.0. Each one was presented in 20% of the trials. To establish these values, the increment of modulation depth estimation proposed by Zwicker and Fastl (2013) (10%) was multiplied by 1, 2, 5 and 10 as well.

Loudness equalization. The loudness of the singletons was equalized as in experiment 1. The levels obtained for each singleton after equalization were 79.5, 79.0, 77.5 and 75.0 dB SPL for the bright singletons with SC of 538, 563, 640 and 768 Hz, respectively, and 80 dB SPL for all the rough singletons. All inter-participants standard deviations of the obtained levels were less than 1 dB SPL.

Procedure. The procedure was the same as in experiment 1, except that the number of trials had to be increased because of the increased number of singletons. Eight blocks of 80 randomly distributed trials each were run for each participant.

Results. The data processing was the same as for experiment 1. For the error rate analysis, 95.0% and 94.6% of the data were kept for experiments 2 and 3, respectively. For the response time analysis, only the data where the participant's response was correct were kept, i.e., 78.6% and 76.4% of the data. The mean response time and error rate across the 20 participants for sequences without singleton were 867 ms (std = 246 ms) and 12.6% (std = 13.1%) for experiment 2, 1058 ms (std = 294 ms) and 15.2% (std = 12%) for experiment 3. The increase in response time for each singleton, i.e., the difference between the condition with the considered singleton and the reference condition without any singleton, is presented in Fig. 2.12 for each value of modulation depth and spectral centroid.

For both experiments 2 and 3, t-tests were conducted with Holm corrections for repeating comparisons. Complete statistics can be found in the Supplementary information (S1 and S2) [3]. Data from experiment 2 confirmed and extended the result of experiment 1 as various bright singletons produced an attentional capture effect. Moreover, the effect increased with SC values : the brighter the singleton, the greater the effect. Experiment 3 showed that roughness is also a feature that triggers an attentional capture effect : the presence of various rough singletons caused significant

3. The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-33496-2. They are reported in appendix 2.

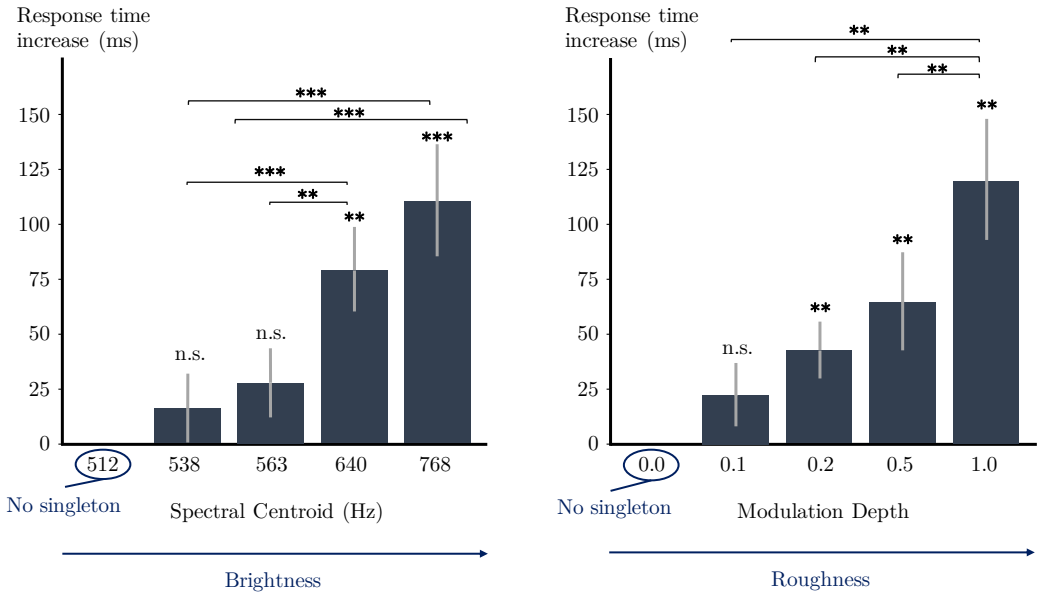

FIGURE 2.12 – Increase in response time (ms) with singleton SC (left, experiment 2) and modulation depth (right, experiment 3). Error bars represent the standard errors across participants in each condition compared to the no-singleton condition. Significances between conditions are displayed on the horizontal braces. * : $p < .05$, ** : $p < .01$, *** : $p < .001$.

behavioral costs. The results confirmed that there is a dependency of salience with the variations of the feature which define the singleton.

Interestingly, the manipulations of the two timbre attributes resulted in comparable effect magnitudes. An increase of a few increments on brightness gives an effect similar to that obtained with an increase of the same number of increments on roughness. This is discussed in the general discussion.

Experiment 4 : symmetrical variations of brightness

Experiment 4 was conducted to study the symmetry or the directionality of the effect. We replicated experiment 2 with SC values for the singleton being either higher or lower than the distractors SC.

Method. *Participants.* 19 participants (8 females, 11 males) took part in the experiment 4. They ranged in age from 18 to 32 years (mean age : 25 ± 4 years). They were all consenting and reported normal hearing. An audiometry in the frequency range between 0.125 and 8 kHz was performed for each participant

and revealed no hearing impairment. Participants gave written informed consent and received financial compensation for their participation.

Apparatus. The apparatus was the same as in the first experiment, except that it took place in the INSEAD-Sorbonne Université Behavioural Lab, in soundproofed rooms.

Stimuli. The distractor and target SC was equal to 631 Hz. The singleton SC was 2 and 4 jnd higher or lower than the distractor one, i.e., 512, 569, 696 or 768 Hz. Each one was presented in 20% of the trials. All the sounds were equalized in loudness (12 participants with the same procedure as in experiment 1) : the obtained levels were 80, 79, 77 and 75 dB SPL for the singletons with SC at 512, 569, 696 and 768 Hz respectively, and 78 dB SPL for the distractor. All inter-participants standard deviations were less than 1 dB SPL.

Results. The data processing was the same as for experiment 1. For the error rate analysis, 94.7% of the data were kept. For the response time analysis, only the data where the participant's response was correct were kept, i.e., 87.1% of the data. The mean response time and error rate across the 19 participants for sequences without singleton were 940 ms (std = 195 ms) and 5.1% (std = ± 8.4%). The increase in response time for each singleton, i.e., the difference between the condition with the considered singleton and the reference condition without any singleton, is presented in Fig. 2.13. Complete statistics can be found in the Supplementary information (S3).

The effect magnitudes are comparable to those obtained in experiment 2. A clear symmetry is observed in experiment 4 : the effect of a brighter singleton is the same as the one of a less bright singleton, if both vary absolutely by the same amount of perceived brightness. This result tells us that it is the absolute variation of the singleton feature that modulates the attention capture. The results of experiments 1, 2, 3 and 4 can be summarized in Fig. 2.14, which shows the driving of response time increases by the perceived variations in the singleton feature. These perceived variations are shown in terms of jnd values.

Interestingly, a linear relationship seems to emerge between increases of perceived brightness (combined across experiment 1, 2 and the positive variations in experiment 4) and response time increase ($r_{Pearson}(3) = 0.99$, $p < 0.001$, slope =

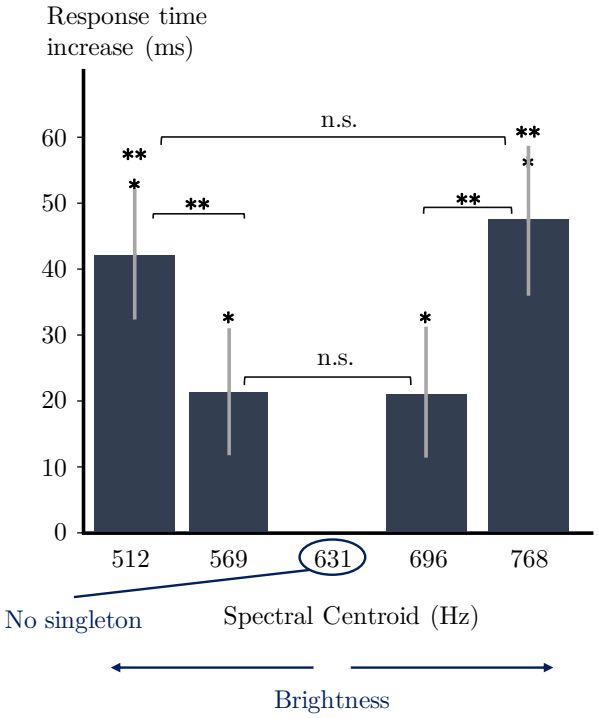

FIGURE 2.13 – Increase in response time (ms) with singleton SC (experiment 4). Error bars represent the standard errors across participants in each condition compared to the no-singleton condition. Significances between conditions are displayed on the horizontal braces. * : p < .05, ** : p < .01, *** : p < .001.

14.0 ms — std error = 0.9 ms), and for perceived roughness as well ($r_{Pearson}(3) = 0.99$, $p < 0.01$, slope = 12.4 ms — std error = 0.9 ms). This relationship is only valid for this range of feature variations and is discussed in the general discussion.

Experiment 5 : combination of roughness and brightness

Experiment 5 was conducted to study the additivity of the effects of different features variations. We replicated experiment 2 with four different singletons, having different combinations of roughness and brightness. The singleton could have two different SC combined with two different amplitude modulation depths.

Method. *Participants.* Nineteen participants (9 females, 10 males) took part in the experiment 4, whose ages ranged from 21 to 36 years (mean age : 26 ± 5 years). They were all consenting and reported normal hearing. An audiometry in the frequency range between 0.125 and 8 kHz was performed for each participant

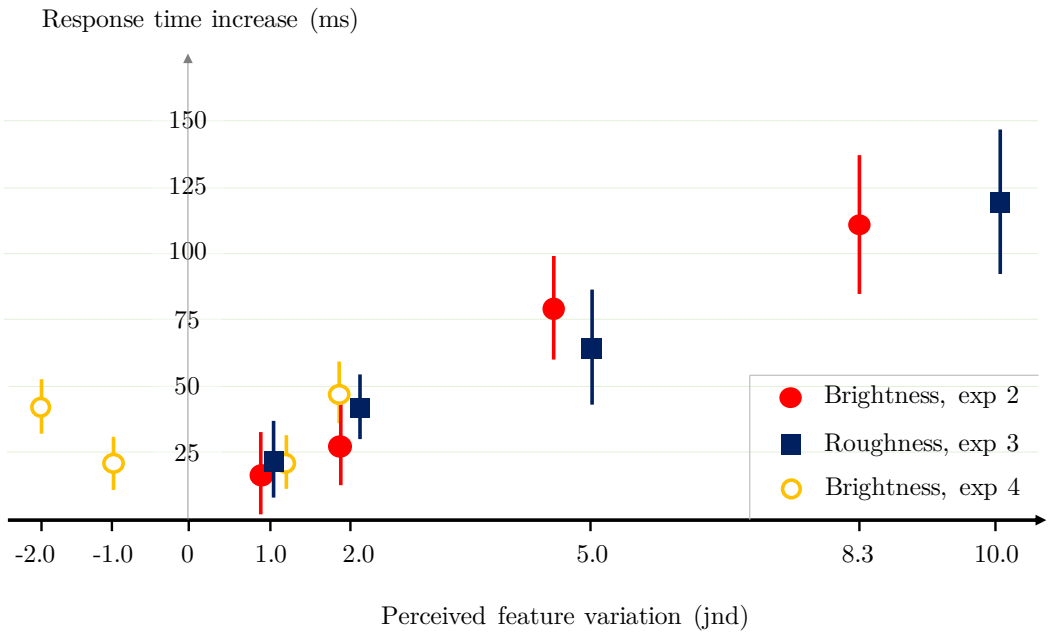

FIGURE 2.14 – Increase in response time (ms) depending on the singleton perceived feature variations (jnd) in experiments 2, 3, and 4. Error bars represent the standard errors across participants in each condition compared to the no-singleton condition.

and revealed no hearing impairment. Participants gave written informed consent and received financial compensation for their participation.

Apparatus. The apparatus was the same as in the first experiment, except that it took place in the INSEAD-Sorbonne Université Behavioural Lab, in soundproofed rooms.

Stimuli. The distractor and target SC was equal to 512 Hz, and they were not modulated, i.e., null roughness. The singleton SC was 2 or 5 jnd higher than the distractor one, i.e., 564 and 653 Hz. The singleton modulation depth was 2 or 5 jnd higher as well, i.e., 0.2 and 0.5. The four singletons were thus obtained with the four combinations of these SC and modulation depths. Each one was presented in 20% of the trials. All the sounds were equalized in loudness (12 participants with the same procedure as the one used in experiment 1) : the obtained levels were 79 dB SPL for the singletons with 2 jnds of brightness, 77.5 dB SPL for the singletons with 5 jnds of

brightness. All inter-participants standard deviations were less than 1 dB SPL.

Results. The data processing was the same as for experiment 1. For the error rate analysis, 94.9% of the data were kept. For the response time analysis, only the data where the participant's response was correct were kept, i.e., 74.9% of the data. The mean response time and error rate across the 19 participants for sequences without singleton were 994 ms (std = 158 ms) and 17.4% (std = 12.7%). The increase in response time for each singleton, i.e., the difference between the condition with the considered singleton and the reference condition without any singleton, is presented in Fig. 2.15. Complete statistics can be found in the Supplementary information (S4).

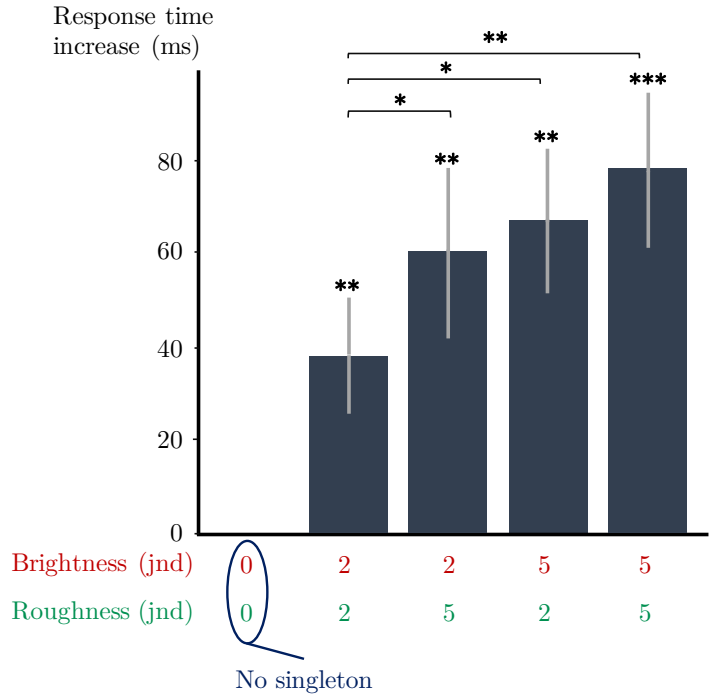

FIGURE 2.15 – Increase in response time (ms) with singleton SC and modulation depth (experiment 5). Error bars represent the standard errors across participants in each condition compared to the no-singleton condition. Significances between conditions are displayed on the horizontal braces. * : $p < .05$, ** : $p < .01$, *** : $p < .001$.

The effect produced by a 2 + 2-jnds variation here is comparable to that produced by a 2-jnds variation in experiments 2 and 3. It is uncertain whether this is due to a non-additivity of the effects of the combined features or whether participants

were simply less subject to attentional capture in this experiment. Nevertheless, within their range of magnitudes, the response times in experiment 5 appear to increase linearly with the addition of the perceptual variations on the two dimensions ($r_{Pearson}(3) = 0.99$, $p < 0.01$, slope = 8.5 ms — std error = 0.4 ms). In other words, the effect seems to be additive across dimensions in this range of values.

Public significance statement. These findings provide evidence that the perception of certain auditory features drives the ability of sounds to capture our attention, according to laws that are revealed.

General discussion

Results from experiment 1 showed that a singleton defined by its timbre, specifically its brightness, captured participants' attention despite being irrelevant to the task they had to perform. Experiment 2 proved that the effect magnitude was driven by the singleton brightness. Experiment 3 showed that a different attribute, roughness, also drives the attentional capture effect. Results from experiment 4 and 5 revealed that this effect is symmetrical, i.e., that only the absolute perceived deviation matters, and additive, i.e., that combining features produces the addition of the effects that each feature variation produces alone.

Thus, in a series of 4 different experiments (2, 3, 4 and 5), a driving of attentional capture by the singleton feature was observed. All else being equal in the experiments, the participants' attentional state remained identical across the different values of the singleton feature. Nevertheless, the magnitude of the effect increased with increasing brightness or roughness variation. The results cannot be explained by increasing singleton-target similarity, because the timbre variations defining the singletons did not make them more similar to the target. Since the increased response times cannot be explained by top-down processes that change with the value of the singleton feature, the observed relationships represent purely feature-driven components of the effect. In other words, the bottom-up component of the attentional capture effect is revealed here, not only confirming its existence (Theeuwes, 1993), but also revealing its pattern.

Thus, by varying the timbre of the tones while keeping the participants' attentional state fixed, we were able to elicit only the bottom-up component of attentional capture. However, the nature of our protocol itself could raise questions about the participants' attentional state and thus the origin of the capture. The contingency

on participants' attentional state (Folk et al., 1992) is questionable here. Indeed, according to the contingency hypothesis (Luck et al., 2021), the task leads to an attentional state that favors the detection of singletons, and this is why attention is captured by the singleton. However, in the present experiments, there were two single items (out of five) in 80% of the trials, and the singleton was one out of 4 possible singletons. Furthermore, all sounds had a fundamental frequency randomly drawn from a broad uniform distribution of 20 Hz. Thus, the variability of the items was increased in our protocol, and the target was not a single item among all identical items. The single-item detection strategy may therefore no longer be advantageous in this setting, and the adaptation of the singleton to participants' attentional state may be different from that which was traditionally thought to be responsible for detection in this paradigm. Further work is needed to understand the interactions between the bottom-up component revealed here and top-down processes, and to address the issue of the compatibility of these results with the contingent capture approach. For example, it would be important to investigate how the driving by the singleton features evolves as participants change their attentional state.

The feature-driven relationships obtained make it possible to observe and compare how different features modulate attention capture. Indeed, the marginal increase of the effect (the derivative of the curves of response time increases with the perceptual variations of the feature) can be interpreted as the weight of the feature in the sound salience. Interestingly, in experiments 2 and 3, both features drove the effect in a similar way. Either these two features are by chance equally responsible for the salience of a sound, or it is the perceived deviation on each dimension that is important in making a sound salient. This evolution of attentional capture with variations of different features therefore deserves to be confirmed through more experiments involving more features (harmonicity, attack time, spectral flux...). If a similar driving is found for other features, it would show that it is precisely how different the sound is perceived that matters to trigger attentional capture, regardless of the feature used. On the contrary, some features could drive the effect with more or less power. This would lead to a hierarchy of features that influence the salience of a stimulus in terms of its ability to capture attention.

Furthermore, the combined results of experiments 1, 2, 3 and 4 (summarized in Fig. 2.14) reveal a monotonic relationship between the perceived difference of the singleton feature (quantified in just-noticeable differences) and the increase in reaction time. Thus, the attentional capture effect increases progressively with the

perceived difference, according to a law that appears to be linear in the range of deviations tested. This law cannot extend over a very wide range of values, as the capture effect must saturate at some point. In any case, we observe that there is no threshold effect, the function is monotonic and continuous. A more precise and extensive determination of this function could also be further investigated in future studies.

This work also brings new insights into the understanding of auditory salience itself, confirming the importance of timbre in this property. Both brightness and roughness were found to be responsible for an attentional capture by irrelevant sounds. It therefore appears that timbre is also a key dimension in directing auditory attention, in addition to the main dimensions of frequency and intensity highlighted by Dalton and Lavie (2004). The results on brightness confirm the findings that previously led some researchers to consider this feature in their salience model (Huang et al., 2017 ; Tordini et al., 2013). Roughness has only recently been included in some form : Kothinti et al. (2021), for example, added average fast temporal modulations to the latest version of their model. The relationship found between attentional capture and feature variations seems to be supported by both features and deserves further investigation, either in other contexts (other tasks, more complex environments...) or with other features.

Our results show that attention capture is driven by absolute deviations of the sound features. In other words, the features do not have an intrinsic polarity with respect to salience (e.g., the brighter, the more salient). Rather, it is a dissimilarity effect that modulates it. This is consistent with predictive coding and theories of auditory deviance detection (Winkler, 2007). They suggest that the deviations between the prediction and what is subsequently perceived determine auditory salience and trigger notified events (Kaya and Elhilali, 2014 ; Southwell et al., 2017). Here, we support these theories by showing that absolute deviations of the sound features directly modulate the magnitude of the attentional capture effect, i.e., their salience.

Finally, our findings are interesting from the perspective of auditory salience modelling, which could be improved by knowing the relevant parameters to consider and how salience depends on their variations. The approach taken so far is to consider the absolute and normalized feature variations over time (Huang et al., 2017 ; Kaya and Elhilali, 2014,1), without implying a more elaborate modulation of attention with these variations. The additivity of the effect produced by different feature variations provides insights into how to combine them (Kaya and Elhilali,

2017). An interesting avenue might be to consider more complex interactions and to go deeper in the understanding of the mechanisms underlying auditory salience.

Conclusion

This work provides contributions on a theoretical, methodological and practical level. From a theoretical point of view, a driving of attention capture by a stimulus feature was revealed. This modulation of bottom-up attention was found to be monotonic and similar for the two timbre attributes studied here : brightness and roughness. The experiment with variations in brightness highlighted symmetric properties, and the experiment with combinations of both attributes underlined the additive character. Methodologically, a way to measure the feature-driven component of attention was proposed : it implies modulating the singleton features in an additional singleton paradigm while keeping the attentional state constant. From a practical perspective, the results may enrich salience models that can include these features and the way they modulate salience in their implementation.

Finally, this study opens perspectives and calls for further studies. The extendibility of the modulation law to more features and to a wider range of feature variations, its dependence on attentional sets and top-down processes, and a higher resolution of the modulation curves deserve further investigation.

Data availability

All data are available at https://github.com/BouvierBaptiste/Revealing-the-stimulus-driven-component-of-attention-through-modulations-of-auditory-salience-by-tim.git.

2.4 Conclusion

Nous avons ainsi mis en évidence la composante *stimulus-driven* de l'attention. Autrement dit, nous avons montré que des sons peuvent s'imposer à nous par le biais de certaines caractéristiques acoustiques/psychoacoustiques, selon des lois qui semblent ici émerger : *linéarité* de la relation entre sensation auditive et temps de capture attentionnelle, *symétrie* des effets induits par des variations d'attributs dans des directions opposées et *additivité* des effets induits par des variations d'attributs différents. Nous confirmons ainsi l'importance de propriétés du timbre dans la modulation de l'attention auditive, et validons, sur un plan psychophysique, les hypothèses sur lesquelles se fondent les modèles de prédiction de saillance auditive.

D'un point de vue méthodologique, le principe expérimental proposé dans cette étude permet d'isoler et d'observer l'effet ascendant de propriétés sonores sur l'attention auditive. Ce principe pourrait ainsi être repris et adapté pour explorer toute une variété de propriétés sonores et comparer leurs influences respectives. La similarité des effets produits par les variations de rugosité et de brillance observée dans cette étude invite d'ailleurs à explorer cette question.

Par ailleurs, les manipulations sonores réalisées dans cette étude pourraient être reprises pour étudier l'effet de la capture attentionnelle dans d'autres paradigmes, mettant en jeu d'autres états attentionnels avec d'autres tâches. Nous avons en effet pour le moment observé l'effet de la saillance dans un paradigme mettant en jeu des séquences sonores simples (des successions brèves de cinq sons similaires). Nous souhaiterions observer comment l'effet de ce type de manipulation peut affecter la perception de séquences plus complexes.

Or, nous avons constaté au chapitre 1 que les principes d'analyse de scènes auditives pouvaient être affectés par des mécanismes ascendants (cf. 1.3.1). Entre autres, le phénomène de primauté du traitement holistique a été mis en

évidence dans la perception de scènes impliquant un traitement à différents niveaux (cf. 1.3.2). L'influence d'un mécanisme ascendant tel que la saillance sur cet effet n'a, à notre connaissance, été exploré que dans la modalité visuelle. Il pourrait être intéressant d'observer si une manipulation de saillance similaire à l'une de celles réalisées dans ce chapitre peut affecter un effet comme celui de primauté du traitement holistique dans des séquences sonores plus complexes. C'est l'objet du chapitre 3.

Chapitre 3

Saillance et primauté du traitement holistique de l'information

"Le bonheur vient de l'attention prêtée aux petites choses, et le malheur de la négligence des petites choses."

Proverbe chinois

3.1 Introduction

Les résultats du chapitre précédent ont permis de confirmer que certaines propriétés sonores sont susceptibles de moduler la saillance auditive. Cependant, le phénomène de capture attentionnelle que nous avons mis en évidence se produisait au sein de séquences sonores très simples, dans lesquelles la tâche se limitait à la détection d'une cible parmi cinq sons identiques. Nous ne savons pas comment ce phénomène affecte la perception de scènes qui impliquent des mécanismes perceptifs et un traitement cognitif plus complexes. Nous avons vu en section 1.3.1 que la perception de scènes complexes est soumise à une hiérarchie entre traitement au niveau global et au niveau local. Ce type de phénomène pourrait-il être affecté par la saillance auditive ?

La première question abordée dans ce chapitre concerne l'effet de la saillance de sources sonores sur l'organisation du traitement de l'information sonore. Nous nous demandons en particulier si la saillance a un effet sur l'analyse temporelle d'une scène sonore, au niveau local et au niveau global. En d'autres termes, est-ce que la saillance peut être associée à une réorganisation perceptive favorisant l'information locale plutôt que l'information globale ? Est-ce que la saillance a un effet sur la perception globale de la scène sonore ? Pour ce faire, le timbre sera manipulé en s'inspirant des résultats du chapitre précédent.

La deuxième question concerne la manière dont cet effet potentiel pourrait être modulé par des processus cognitifs descendants (voir 1.3.1.1), ici associés à l'expertise musicale. En d'autres termes, est-ce que la saillance peut être associée à une réorganisation perceptive indépendamment des mécanismes cognitifs descendants impliqués ? Pour ce faire, l'expertise musicale des participants sera un facteur contrôlé.

Ces travaux ont fait l'objet d'une collaboration avec Emmanuel Ponsot (chercheur de l'équipe Perception et Design Sonores de l'IRCAM) pour l'adaptation du paradigme issu des travaux précédents avec Patrick Susini. Ils ont été présentés au Forum Acusticum 2023 (Bouvier et al., 2023c). Comme

au chapitre 2, notre question de recherche implique la mise en oeuvre d'un paradigme permettant de réaliser une mesure expérimentale appropriée au phénomène. Une fois de plus, c'est dans la modalité visuelle que les travaux expérimentaux ont été initiés.

3.1.1 Hiérarchie du traitement local/global de l'information visuelle

3.1.1.1 Mise en évidence expérimentale

Navon (1977) est le premier à s'être intéressé de près à la hiérarchie du traitement local/global de l'information visuelle. Dans la troisième expérience de son étude, de grands caractères constitués de plus petits caractères étaient présentées à des participants qui devaient identifier soit le grand caractère, soit les petits. Plus précisément, les participants pouvaient voir apparaître un grand H, un grand S, ou un grand rectangle. Ces caractères globaux étaient constitués de caractères locaux plus petits : des H, des S ou des rectangles (voir figure 3.1).

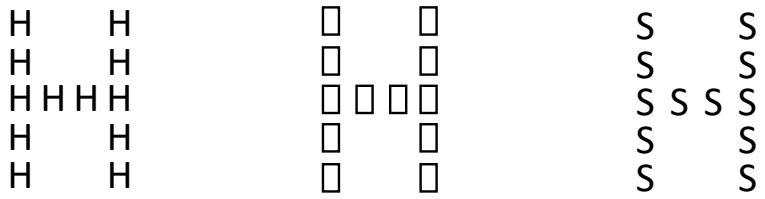

FIGURE 3.1 – Caractères utilisés dans le paradigme local/global de Navon (1977). À gauche, les informations au niveau local et global sont congruentes, au centre neutres, à droite, incongruentes.

Dans la tâche globale, les participants devaient indiquer si le caractère global était un H ou un S. Dans la tâche locale, ils devaient indiquer si les caractères locaux étaient un H ou un S. Navon a mesuré les temps de réponse dans les deux tâches en fonction des conditions, congruente ou incongruente entre les caractères au niveau local et ceux au niveau global. Ses résultats

sont présentés en figure 3.2.

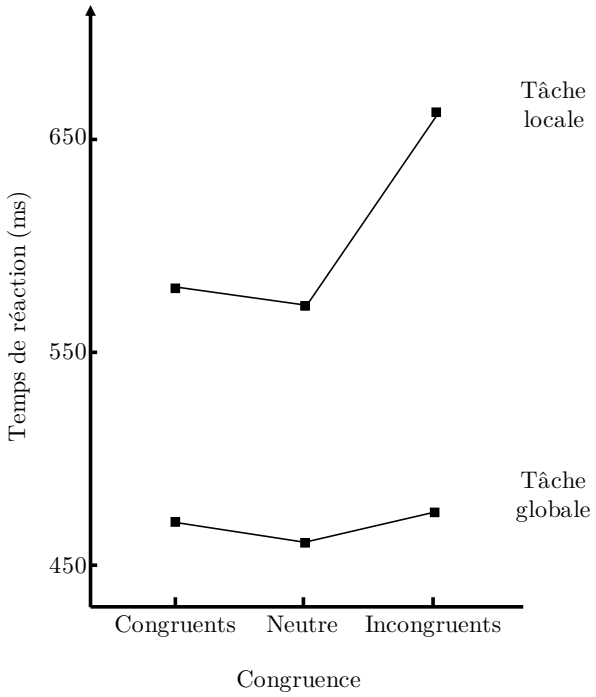

FIGURE 3.2 – Temps de réaction en fonction de la tâche et de la congruence entre les caractères locaux et globaux dans la troisième expérience de Navon (1977).

Les résultats montrent que les participants répondent plus vite pour identifier le caractère global. De plus, alors que les caractères au niveau local ne semblent pas perturber la reconnaissance du caractère au niveau global, l'inverse n'est pas vrai : les participants ne peuvent pas ignorer le caractère global lorsqu'il essaient de se focaliser sur les caractères locaux. Dans une quatrième expérience, Navon a montré que dans une tâche de comparaison de formes géométriques par paires, les participants détectaient plus souvent les différences globales que locales.

Il semble donc que le traitement de l'information au niveau global soit réalisé en priorité et ne puisse pas être outrepassé pour traiter les informations au niveau local. Il y a donc bien un effet de précédence globale, ou une

primauté du traitement holistique de l'information visuelle (voir section 1.3.2).

3.1.1.2 Effet de la saillance sur la hiérarchie du traitement local/global de l'information visuelle

L'effet de précédence globale obtenu dans la modalité visuelle et présenté en section 3.1.1 peut être modulé dans certaines conditions. Mevorach et al. (2006) a montré que la saillance des stimuli pouvait affecter l'organisation hiérarchique du traitement local/global, dans une adaptation du paradigme de Navon (1977).

Les caractères présentés aux participants pouvaient être manipulés pour rendre la lettre globale ou les lettres locales plus saillantes. Les stimuli sont présentés en figure 3.4.

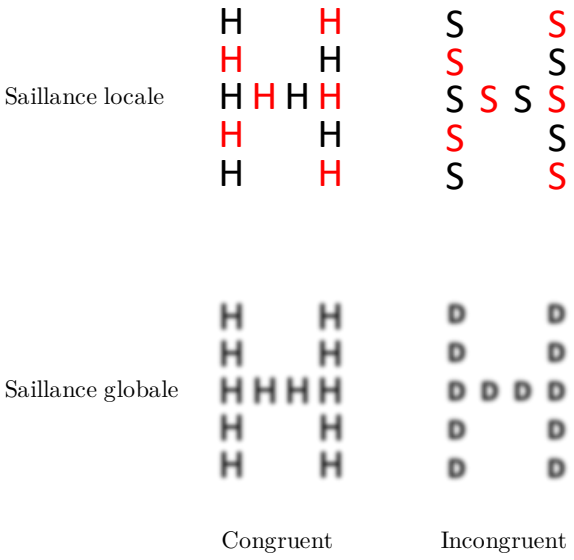

FIGURE 3.3 – Caractères utilisés dans le paradigme local/global de Mevorach et al. (2006). À gauche, les informations aux niveaux local et global sont congruentes, à droite, incongruentes. En haut, la saillance est portée sur le niveau local (alternance de couleur localement), en bas sur le niveau global (caractères locaux uniformes et floutés).

Les temps de réaction sont mesurés et présentés en figure 3.3. On note que lorsque la saillance est portée sur le niveau local, les participants deviennent plus performants pour détecter les modifications locales. De même, ils détectent mieux les modifications globales lorsque la saillance est portée à ce niveau. Dans ces deux cas, la congruence entre les informations aux deux niveaux n'a plus d'importance : les performances sont les mêmes qu'elles soient congruentes ou incongruentes.

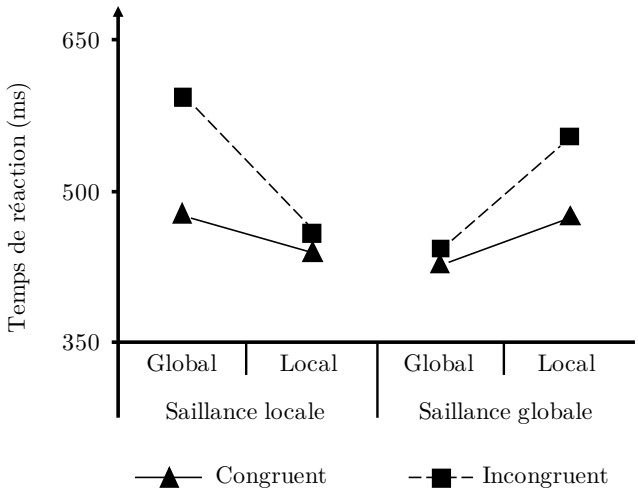

FIGURE 3.4 – Temps de réaction mesurés dans l'expérience de Mevorach et al. (2006). À gauche, les résultats lorsque la saillance est portée sur le niveau local, à droite sur le niveau global, à chaque fois dans la tâche globale ou la tâche locale, en fonction de la congruence entre les informations locales et globales.

Cette étude montre que la présence de saillance dans les stimuli oriente l'attention sur les informations au niveau favorisé. Autrement dit, le mécanisme bottom-up qu'est la saillance influence l'organisation locale/globale du traitement de l'information spatiale visuelle. Il apparaît en effet que ce n'est pas la prévalence de l'information au niveau local ou au niveau global qui importe mais plutôt la saillance de cette information.

3.1.2 Hiérarchie du traitement local/global de l'information auditive

3.1.2.1 Adaptation du paradigme dans la modalité auditive

Comme détaillé en section 1.3.2.2, le paradigme local/global a par la suite été adapté dans la modalité auditive. Justus and List (2005) ont en premier lieu adapté les stimuli à l'audition en passant de la perception visuelle spatiale à la perception auditive temporelle (cf. section 2.2.1 concernant l'équivalence temps-espace entre vision et audition). Ils ont proposé des mélodies de 9 notes successives dont la hauteur pouvait être modifiée au niveau d'une seule note (local) ou d'un groupe de notes (global). Avec ces stimuli, Bouvet et al. (2011) ont montré que les participants étaient plus rapides et plus précis pour détecter les variations globales. D'autres études Black et al. (2017) ; Ouimet et al. (2012) ont ensuite confirmé cet effet de précédence globale dans le traitement de l'information auditive. La dernière adaptation de ce paradigme, s'affranchissant notamment d'un biais pouvant exister dans la mesure du temps de réponse, a été proposée par Susini et al. (2020).

Dans ce paradigme, les participants doivent comparer des paires de mélodies selon leur profil de hauteur, autrement dit la hauteur des notes qui les composent. Ces mélodies, constituées de trois triplets de trois notes chacun, peuvent différer à un niveau local (modification de la mélodie au sein d'un seul des triplets), à un niveau global (transposition de tout un triplet), ou au deux niveaux en même temps (combinaison des deux modifications). Elles peuvent aussi ne pas différer, à l'exception d'une transposition de toute la mélodie. Des exemples de stimuli sont présentés en figure 3.5. Les mélodies cibles peuvent être ascendantes, descendantes, ou suivre des profils ascendant-descendant ou descendant-ascendant.

Dans la tâche locale (ou globale), les participants doivent juger si les deux mélodies sont similaires ou différentes sur le plan local (ou global). Ainsi par exemple, dans la tâche globale, les mélodies de comparaison présentant une modification locale ou pas de modification doivent être jugées comme

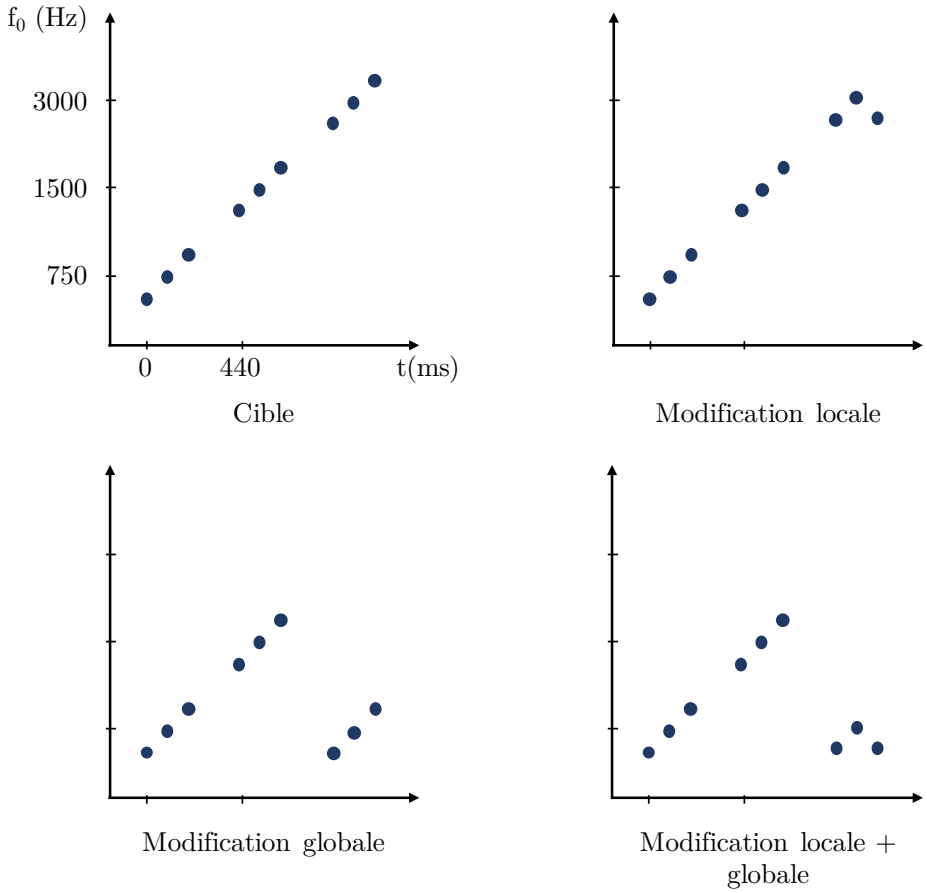

FIGURE 3.5 – Stimuli utilisés dans le paradigme local/global de Susini et al. (2020). La mélodie cible est ici une mélodie ascendante. Les modifications sont proposées dans cet exemple sur le troisième triplet.

globalement similaires. Les mélodies présentant une modification globale ou une modification locale+globale doivent être jugées comme différentes.

Susini et al. (2020) ont mesuré les taux de bonnes réponses dans les différentes tâches et conditions pour en dériver différents indices. D'une part, les indices propres à la théorie de détection du signal (sensibilité d', notamment), d'autre part des indices d'avantage global et d'interférence globale-locale présentés en section 3.3.1.

Leurs résultats confirment les résultats observés dans la modalité visuelle et dans les premières mises en oeuvre de paradigmes local/global dans la modalité auditive. Les participants présentent un avantage global positif : ils détectent plus facilement les modifications au niveau global qu'au niveau local. De plus, les modifications de profil au niveau global perturbent la détection de modifications au niveau local, plus que réciproquement (voir figure 3.6). Autrement dit, comme en vision, le traitement de l'information au niveau global est réalisé en priorité (avantage global) et ne peut être ignoré pour traiter les informations au niveau local (interférence globale-locale).

Cependant, ces résultats dépendent de l'expertise des participants : ils sont valables pour des participants non-musiciens uniquement. En effet, pour des musiciens, il ne semble plus y avoir d'avantage global ni d'interférence globale-locale. Ces tendances semblent même légèrement inversées pour ces derniers. Ils seraient donc capables de traiter l'information au niveau local aussi bien qu'au niveau global, et même d'ignorer les variations globales pour détecter des modifications locales. Ces résultats ont depuis été confirmés (Susini et al., 2023) et précisés : des musiciens amateurs présentent déjà un avantage global réduit voire négatif. Ceci suggère que l'expertise musicale façonne les mécanismes d'écoute analytique, et donne la capacité à orienter son attention sur le niveau de traitement souhaité.

3.1.2.2 Expertise musicale

La pratique musicale permet de développer des capacités améliorées en matière de perception et de traitement de l'information sonore (Herholz and Zatorre, 2012 ; Talamini et al., 2017). Les musiciens sont ainsi meilleurs pour organiser des flux auditifs dans l'espace fréquentiel (Bey and McAdams, 2002,0 ; van Noorden, 1975 ; Wenhart and Altenmüller, 2019) ou dans le temps. Concernant cette dimension, certaines études mentionnées en section 1.3.2.2 (mettant en oeuvre un paradigme d'évaluation du traitement local/global de l'information sonore temporelle) se sont d'ailleurs intéressées à l'influence de l'expertise musicale des participants sur les résultats observés. Elles ont permis de montrer que les musiciens sont plus performants que les

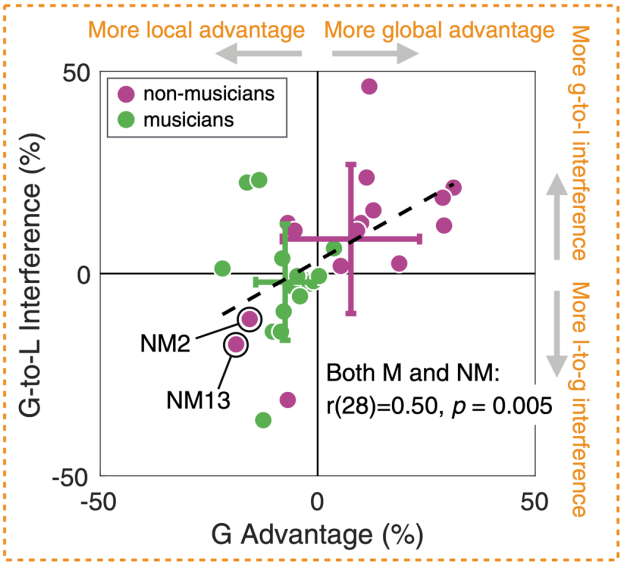

FIGURE 3.6 – Résultats obtenus par Susini et al. (2020). Les performances des participants sont présentées dans le plan (Avantage global, Interférence globale-locale). les non-musiciens présentent un avantage global et une interférence globale-locale positive. Pour les musiciens, ces deux indices sont légérement négatifs.

non-musiciens pour traiter les informations de manière détaillée, c'est-à-dire de se concentrer sur l'information au niveau local plus qu'au niveau global s'il le faut.

Ouimet et al. (2012) et Black et al. (2017) ont ainsi mis en évidence que l'effet de précédence globale était diminué pour les participants experts-musiciens, et ceci parce qu'ils étaient notamment plus performants pour traiter l'information localement.

Ces résultats ont été confirmés et approfondis par Susini et al. (2020) puis Susini et al. (2023) qui ont montré que l'avantage global des non-musiciens était inversé chez les expert-musiciens et les musiciens amateurs.

La pratique musicale semble ainsi permettre de développer des capacités d'écoute analytique renforcées (Bever and Chiarello, 1974), et donc de favoriser la détection des détails dans une mélodie, indépendamment des modifications au niveau global. Ces capacités à pouvoir orienter son attention

au niveau souhaité constituent un exemple de mécanismes top-down mis à l'oeuvre dans le traitement de l'information auditive : le participant musicien a appris à orienter son attention auditive sur l'information de son choix.

3.1.3 Discussion

Ainsi, le paradigme de Susini et al. (2020) permet de sonder une composante particulière du traitement de l'information auditive : la hiérarchie entre traitement au niveau local et traitement au niveau global. Il a notamment confirmé que l'information était traitée au niveau global de manière privilégiée, et montré que la hiérarchie entre traitement au niveau local et global était réorganisée par l'expertise musicale.
Par ailleurs, s'il a déjà été observé que la saillance a un effet sur l'organisation du traitement de l'information au niveau local et global dans la vision (Mevorach et al., 2006), cette question reste ouverte dans la modalité auditive.

Or, nous nous demandons dans ce chapitre comment la saillance peut affecter le traitement local/global de l'information auditive, et comment des mécanismes cognitifs descendants sous-tendus par l'expertise musicale peuvent interagir avec cet effet (voir section 3.1).

Pour répondre à cette question, nous avons donc choisi de mettre en oeuvre le paradigme de Susini et al. (2020) et d'y observer l'influence de la présence de sons saillants, tout en contrôlant le niveau d'expertise musicale des participants.

Les résultats du chapitre précédent ayant montré que des manipulations de brillance (associée au centre de gravité spectral) pouvaient rendre des sons saillants au sein d'une séquence sonore, nous avons utilisé le même type de manipulation dans ce paradigme.

3.2 Expérience

3.2.1 Participants

Vingt participants ont pris part à l'expérience : 13 non-musiciens (4 femmes, âge moyen : 31,4 ± 11,0 ans) et 7 experts-musiciens (1 femme, âge moyen : 41,4 ± 15,9 ans). Aucun n'a déclaré avoir des problèmes d'audition. Ils ont donné leur consentement par écrit avant l'expérience et ont été rémunérés pour leur participation. Les critères d'expertise musicale étaient les suivants : les musiciens experts étaient des personnes ayant une solide formation musicale dans des institutions françaises telles que les Conservatoires Nationals à Rayonnement Régional (CRR), se considérant comme des musiciens, ayant une pratique quotidienne, plus de six ans d'apprentissage musical théorique et instrumental, et jouant avec d'autres musiciens dans des groupes ou des ensembles orchestraux. L'expertise musicale a été évaluée par le questionnaire Gold-MSI (Müllensiefen et al., 2014), dont les détails sont rapportés en annexe 3. Le groupe des non-musiciens était composé de volontaires se déclarant non-musiciens, c'est-à-dire n'ayant aucune formation musicale.

3.2.2 Équipement

Les sons ont été présentés aux auditeurs par l'intermédiaire d'un casque Beyerdynamic DT-770 PRO et d'une carte son Focusrite Scarlett 2i2 à un niveau de sortie de 70 dB SPL. Le niveau sonore a été mesuré en amont à l'aide d'un sonomètre de type 2250-S de Bruel & Kjær. L'expérience s'est déroulée dans une cabine insonorisée à double paroi d'Industrial Acoustics Company (IAC). L'interface de test a été codée dans l'environnement Max MSP (v8) [1] sur un ordinateur Apple Mac mini.

1. https://cycling74.com/products/max

3.2.3 Stimuli

La construction des stimuli est inspirée de Susini et al. (2020) et nous en reprécisons les détails ici. Chaque stimulus est composé de 9 notes, segmentées en trois triplets de trois notes. Le niveau local est défini comme la structure de hauteur à l'intérieur des triplets, et le niveau global comme la structure de hauteur formée par la hauteur moyenne des trois triplets. À chaque essai, les stimuli sont présentés par paire aux participants, une cible et une comparaison, séparées par un silence de 500 ms.

La durée des notes est de 100 ms, les intervalles entre les notes au sein de chaque triplet sont de 10 ms et les intervalles entre les triplets de 120 ms, ce qui donne des séquences de 1 200 ms. Le centre de gravité (moyenne sur une échelle de log-fréquence) de la hauteur du deuxième triplet est choisi selon une distribution uniforme aléatoire ([400 ; 1000] Hz). Cette plage de hauteur est plus restreinte que celle de Susini et al. (2020) car les notes sont enrichies d'harmoniques dans le présent protocole, et les plus aiguës d'entre elles auraient pu être perçues comme trop stridentes.

Les séquences sont ensuite structurées de manière à respecter des intervalles musicaux spécifiques : il y a toujours une différence de 4 demi-tons entre deux notes au sein d'un triplet, et il y a toujours une différence d'une octave entre les centres de gravité des hauteurs de deux triplets consécutifs. Cette construction des stimuli sur une échelle musicale est un facteur déterminant pour mettre en avant les performances des experts-musiciens, habitués à ce type d'intervalles.

Notes

Toutes les notes suivent la structure harmonique de Bouvier et al. (2023b), présentée en section 2.3 : chaque note avec une fréquence fondamentale f_0 a n harmoniques (n dans [1, 20]), la nième harmonique f_n ayant une fréquence $n * f_0$ et un poids $1/n^\alpha$. Ainsi, la variation de α modifie le centre de gravité spectral (CGS) du son, et donc sa brillance perçue. Pour les notes ordinaires, $\alpha = 5$. Pour les notes dites brillantes, $\alpha = 1{,}5$. Les niveaux des

notes sont normalisés en sonie sur l'ensemble des fréquences à l'aide de la courbe d'isosonie ISO226 à 70 dB SPL25.

Stimuli cible

Nous avons utilisé des profils de hauteur monotones ascendants ou descendants (respectivement [A] et [D] dans la suite du chapitre) tels qu'employés, entre autres, dans des études antérieures (Bouvet et al., 2011 ; Justus and List, 2005 ; Ouimet et al., 2012). Seuls les profils ascendants et descendants sont retenus dans notre version du protocole, car les différentes conditions de saillance multiplient le nombre de facteurs à croiser et donc de stimuli à présenter aux participants. L'expérience étant déjà conséquente en terme de durée, nous avons donc restreint le facteur profil à ces deux modalités. Chaque triplet est indiqué par C_j, correspondant au centre de gravité des fréquences fondamentales des trois sons au sein d'un triplet, j indiquant sa position dans la séquence, de 1 à 3. Pour chaque stimulus cible, la valeur de C_2 est d'abord choisie aléatoirement au sein une distribution uniforme entre 400 et 1 000 Hz. Ensuite, les valeurs de C_1 et C_3 sont placées à ± 1 octave de C_2.

Stimuli de comparaison

Pour les stimuli de comparaison, C_2 est également choisi sur une distribution aléatoire uniforme entre 400 et 1 000 Hz. Il y a donc toujours au moins une transposition globale de hauteur entre la cible et les stimuli de comparaison. Quatre types de stimuli de comparaison peuvent être présentés :

— aucune modification, à l'exception de la transposition globale (No) ;
— une modification locale (L) ;
— une modification globale (G) ;
— une modification à la fois locale et globale (L + G) ;

Une modification locale correspond à l'altération du profil de hauteur à l'intérieur d'un triplet (transposition d'une seule note à l'intérieur du triplet

modifié). Une modification globale consiste à modifier le profil de hauteur global (transposition d'un triplet entier).

Lorsque des modifications locales et globales sont appliquées simultanément (condition L + G), elles se produisent sur le même triplet. Chaque modification peut se produire sur le premier ou le troisième triplet avec la même probabilité.

Conditions de saillance

À chaque essai, la paire de stimuli présentée peut être affectée ou non par une manipulation de saillance : $2/3$ des essais contiennent une manipulation de saillance. On distingue ainsi différentes conditions, chacune dans $1/3$ des essais :

— la condition de *saillance nulle* ;
— la condition de *saillance congruente* : les notes du triplet modifié sont rendues saillantes ;
— la condition de *saillance incongruente* : un des deux autres triplets (avec autant de probabilité pour chacun) est rendu saillant ;

Un exemple de stimulus est présenté en figure 3.7, pour une mélodie ascendante avec une modification locale sur le troisième triplet.

3.2.4 Procédure

Les participants ont pris part à deux tâches distinctes lors de sessions séparées : une session "locale" et une session "globale". Il leur était demandé d'effectuer une tâche de discrimination "similaire/différent" en se concentrant sur le niveau local ou global selon la session. Dans la session locale, ils devaient déterminer si les profils de hauteur des trois triplets étaient similaires ou différents dans le stimulus cible et le stimulus de comparaison, indépendamment du profil global. Dans la session globale, ils devaient déterminer si le profil global (c'est-à-dire l'organisation $C_1\ C_2\ C_3$) était identique ou non, indépendamment des profils locaux de chaque triplet.

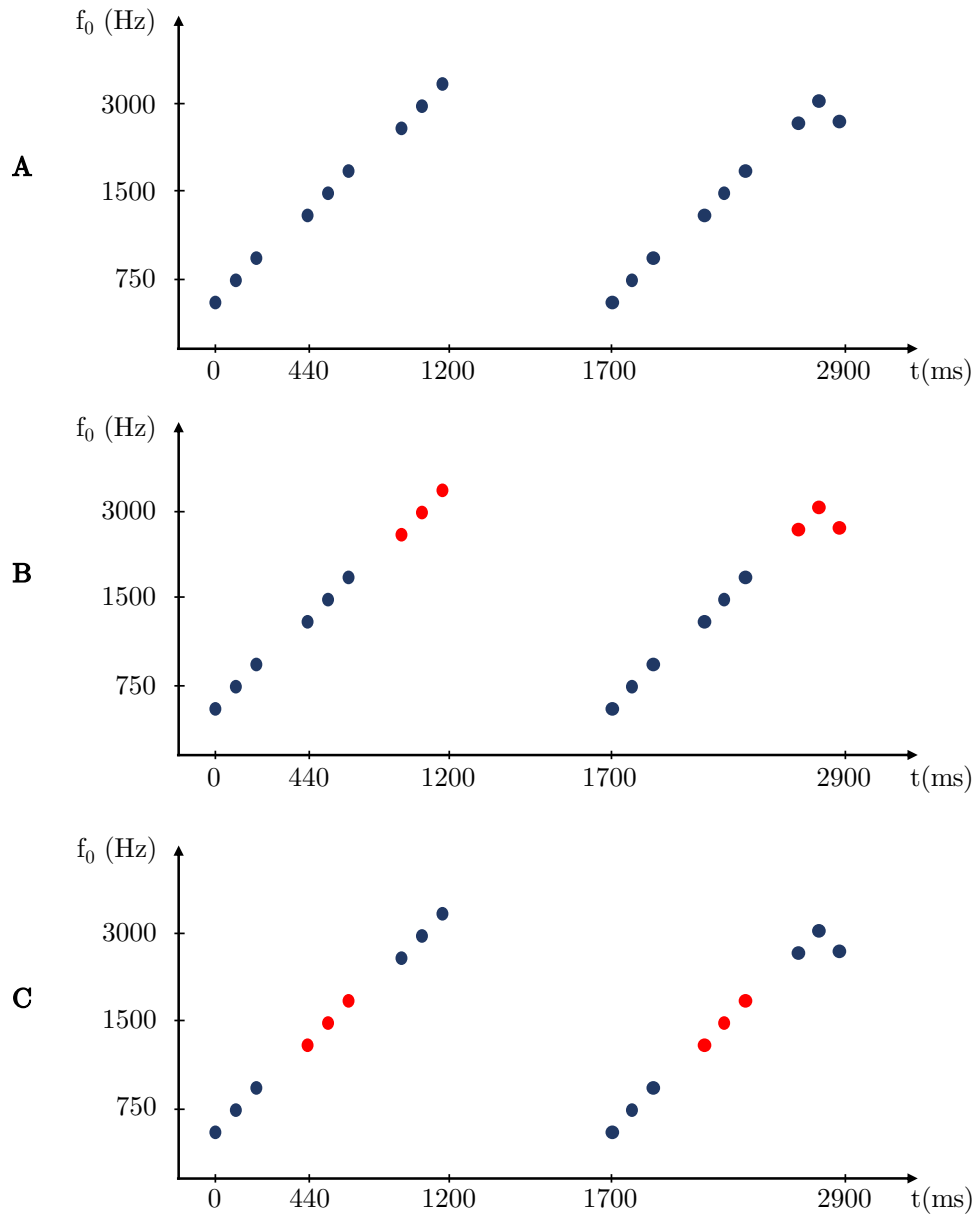

FIGURE 3.7 – Stimuli de l'expérience local/global. Dans cet exemple, la cible suit un profil ascendant, et la modification est une modification locale sur le troisième triplet. La saillance peut être nulle (A), congruente (B) ou incongruente (C) avec la modification de profil. Les notes saillantes sont présentées en rouge.

À la fin de chaque essai, les participants donnaient leur réponse en appuyant sur les boutons "similaire" ou "différent". Ils disposaient d'autant de temps qu'ils le souhaitaient pour répondre. Les participants recevaient un

feedback sur l'exactitude de leur réponse à chaque essai (correct/incorrect). Une fois la réponse donnée, l'essai suivant débutait après un délai de 500 ms. La moitié des participants a commencé par la session locale, l'autre moitié par la session globale.

Compte tenu des quatre variables - deux profils ([A], [D]), quatre conditions de modifications (No, L, G, L + G), deux positions pour la modification de la hauteur (premier ou troisième triplet), trois conditions de saillance (Pas de saillance, Congruent, Incongruent) - il y avait 48 configurations différentes. Afin d'obtenir des scores suffisamment précis, les 48 configurations étaient répétées 10 fois chacune par participant pour chaque session, ce qui donnait un total de 480 essais par session. Tous les essais étaient uniques, puisque la hauteur de la cible était toujours tirée d'une distribution uniforme aléatoire. Chaque session était divisée en 5 blocs de 96 essais et durait environ 1h30. Avant chaque session, les participants se familiarisaient avec les stimuli et la tâche, et effectuaient un bloc d'essais d'entraînement. L'entraînement était validé par l'expérimentateur si les participants obtenaient des résultats supérieurs au hasard (plus de 60% de bonnes réponses).

3.3 Résultats et discussion

3.3.1 Analyse

Les analyses suivent celles proposées par Susini et al. (2020). Cette expérience est basée sur un plan factoriel 2 × [2 × 2 × 4 × 2 × 3] : un facteur inter-participants "Groupe" (Musiciens|Non-musiciens) et cinq facteurs intra-participants "Tâche" (Local|Global) × "Profil" (A|D) × "Condition" (No|L|G|L + G) × "Position" (1|3) × " Saillance " (Nulle|Congruente|Incongruente). Dans le cadre de la TDS, nous avons calculé les matrices de confusion pour dériver les valeurs de sensibilité (d') et de critère de décision (c) pour chaque participant, dans chaque tâche, en fonction des modalités des différents facteurs. Pour chaque tâche, les réponses des participants ont été classées en fonction de la condition :

— Tâche locale : succès = pourcentage de réponses "similaires" dans les conditions No et G ; fausses alarmes = pourcentage de réponses "similaires" dans les conditions L et L + G.

— Tâche globale : succès = pourcentage de réponses "similaires" dans les conditions No et L ; fausses alarmes = pourcentage de réponses "similaires" dans les conditions G et L + G.

Lorsqu'un score était égal à 0 ou 100% dans une condition, le score était remplacé par 1/N ou (N-1)/N pour obtenir la sensibilité et le critère de décision (en accord avec les analyses de Susini et al. (2020)). La sensibilité maximale est donc de 6,2 dans ce cadre.

Pour évaluer plus quantitativement l'avantage global et les effets d'interférence entre les niveaux local et global, deux indices peuvent être calculés (Susini et al., 2020) : le GA ("Global Advantage") et le GL ("Global-to-Local interference"). Si l'on note S_{tc} le score moyen (pourcentage de réponses correctes) d'un participant dans la tâche *t* et la condition *c*, avec l et g se référant aux tâches/conditions locales et globales, alors :

— l'indice de l'avantage global a été calculé comme la différence entre les scores globaux et locaux : $GA = \frac{1}{2}(S_{gl} + S_{gg}) - \frac{1}{2}(S_{ll} + S_{lg})$;
— l'indice d'interférence globale-locale a été calculé comme la différence entre les effets d'interférence globale-locale et d'interférence locale-globale : $GL = (S_{ll} - S_{lg}) - (S_{gg} - S_{gl})$;

3.3.2 Résultats

3.3.2.1 Validation de notre paradigme et confirmation des résultats de la littérature

Distribution des participants

Les premiers résultats concernent la distribution des résultats des participants en fonction de leur groupe d'appartenance : musiciens ou non-musiciens. On peut considérer leurs résultats dans la tâche locale et dans la tâche globale sur tous les essais en condition de saillance nulle et les situer dans le plan (d'_{loc}, d'_{glob}) : en abscisse, leur sensibilité dans la tâche locale, en ordonnée dans la tâche globale (figure 3.8).

On constate une répartition non-homogène avec deux zones privilégiées : l'une dans les plages de valeur de sensibilité plus élevées et occupée majoritairement par les experts-musiciens, l'autre dans les zones de sensibilités faibles occupée par les non-musiciens.

On note que deux non-musiciens ont des performances proches de celles des experts, et un expert des performances proches de celles des non-musiciens. Les participants ont été ré-interrogés à la suite de cette observation (sans qu'ils ne soient informés du résultat). Chez les non-musiciens, tous ont confirmé n'avoir jamais suivi de formation musicale, excepté les participants 9 et 12. L'un a reconnu avoir pratiqué le piano quelques années dans sa jeunesse, l'autre pratiqué le mixage ("DJing") pendant un certain temps. Par ailleurs, l'expert numéro 7 a affirmé s'être formé de manière autodidacte sans suivre de formation musicale, et ses résultats au questionnaire Gold-MSI sont nettement

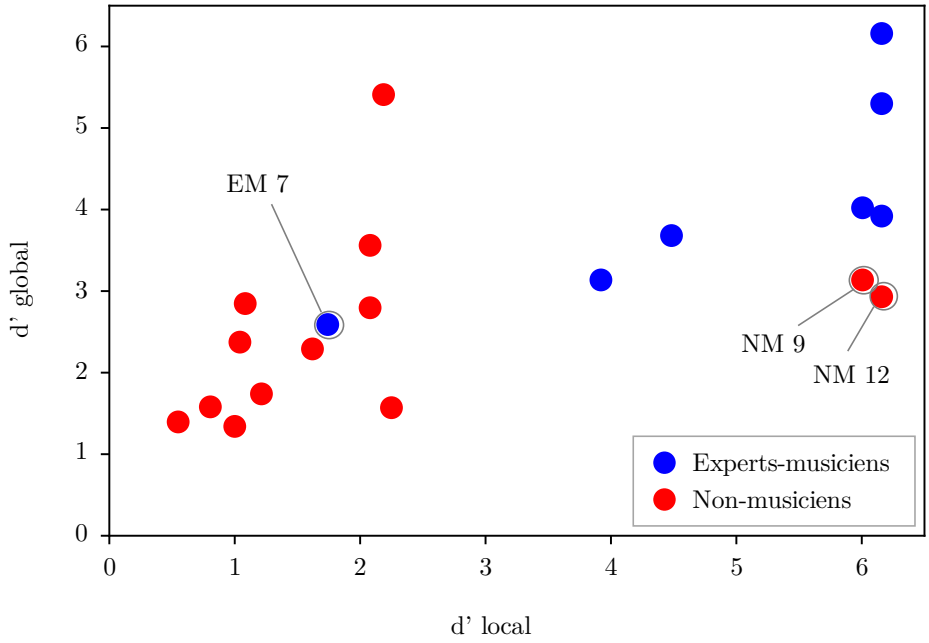

FIGURE 3.8 – Répartition des participants dans le plan $(d'_{local}, d'_{global})$ en condition de saillance nulle. Les musiciens sont en bleu, les non-musiciens en rouge.

inférieurs à ceux des experts-musiciens, notamment pour la partie concernant les capacités d'écoute musicale (voir annexe 3). Or, Susini et al. (2023) ont récemment montré qu'une pratique musicale à niveau amateur suffisait déjà à affecter les performances dans ce paradigme. Afin d'être sûr de ne relever que les effets propres à l'écoute de personnes n'ayant aucune pratique musicale et d'une expertise correspondant aux critères définis en section 3.2.1, nous avons donc retiré les deux non-musiciens 9 et 12, ainsi que l'expert-musicien 7 pour la suite des analyses.

Validation du paradigme

On peut considérer, en figure 3.9, la répartition des participants dans le plan (GA, GL) des indices d'avantage global et d'interférence globale-locale définis plus hauts. Cette représentation ayant été proposée par Susini et al. (2020), elle nous permet de comparer les résultats obtenus dans leur paradigme et dans notre adaptation du paradigme sur les essais en condition de saillance nulle.

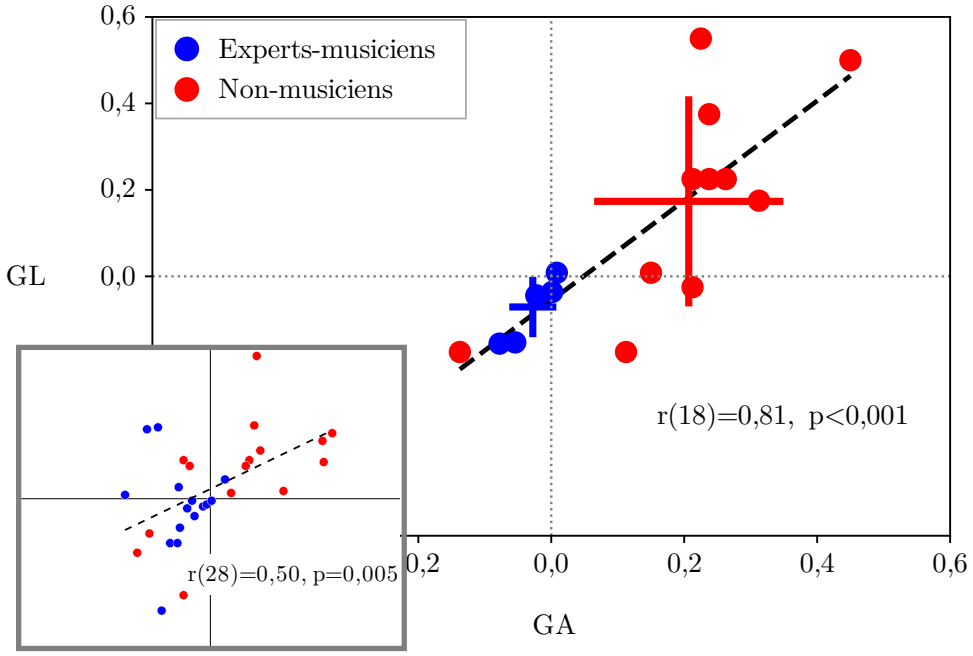

FIGURE 3.9 – Répartition des participants dans le plan (GA, GL). Les musiciens sont en bleu, les non-musiciens en rouge. Barres d'erreur : intervalles de confiance à 95%. En bas à gauche, la même représentation des résultats chez Susini et al. (2020).

On constate encore une fois une répartition hétérogène avec deux zones principales occupées globalement par les participants de chaque groupe. Les musiciens ont un GA et un GL nuls voir légèrement négatifs, les non-musiciens des valeurs plus élevées en moyenne. Ces résultats sont cohérents avec ceux de Susini et al. (2020) : les participants, dans les essais en condition de saillance nulle de notre version du paradigme, se comportent comme les participants sur l'ensemble des essais du paradigme de Susini et al. (2020).

Nous pouvons préciser l'origine de cet avantage global pour les non-musiciens et pas pour les experts-musiciens. Les résultats dans la condition de saillance nulle mettent en effet en lumière que les non-musiciens sont moins performants dans la tâche locale (d' = 1,45) que dans la tâche globale (d' = 2,45). Un test-t a révélé que cet avantage dans la tâche globale par rapport à

la tâche locale était significatif (T(10) = 3,38, p = 0,007, cohen-d = 1,02, power = 0,86). Les experts-musiciens, au contraire, sont meilleurs dans la tâche locale (d' = 5,48) que dans la tâche globale (d' = 4,37). Un test-t a révélé que cet avantage dans la tâche locale par rapport à la tâche globale était également significatif (T(5) = 3,24, p = 0,02, cohen-d = 1,32, power = 0,74).

Les résultats dans la condition de saillance nulle valident ainsi notre paradigme en confirmant les résultats de la littérature (Black et al., 2017 ; Ouimet et al., 2012 ; Susini et al., 2020) : les non-musiciens traitent en priorité l'information au niveau global, et l'expertise musicale semble bien modifier l'organisation du traitement de l'information locale/globale.

3.3.2.2 Effet de la saillance sur la hiérarchie du traitement local/global

Pour observer l'effet de la saillance sur la performance des participants dans les deux tâches, nous présentons la sensibilité dans chaque tâche en fonction des différentes conditions de saillance en figure 3.10 (les trois participants exclus en partie précédente le sont toujours dans ces analyses).

Sur cette figure, on visualise en premier lieu que les musiciens sont plus performants que les non-musiciens quelle que soit la condition de saillance. On y retrouve également les résultats mentionnés plus haut, dans la condition de saillance nulle : les non-musiciens sont plus performants dans la tâche globale que dans la tâche locale, inversement pour les experts-musiciens.

On observe ensuite que ces résultats sont modifiés dans les autres conditions de saillance, de manière différenciée pour les non-musiciens et les experts-musiciens. Des tests-t ont été réalisés dans chacune de ces conditions comparée à la condition de saillance nulle. Les p-values issues des multiples tests ont ensuite subi une correction de Benjamini-Hochberg (dite du "False Discovery Rate") et sont notées p_{corr} dans la suite.

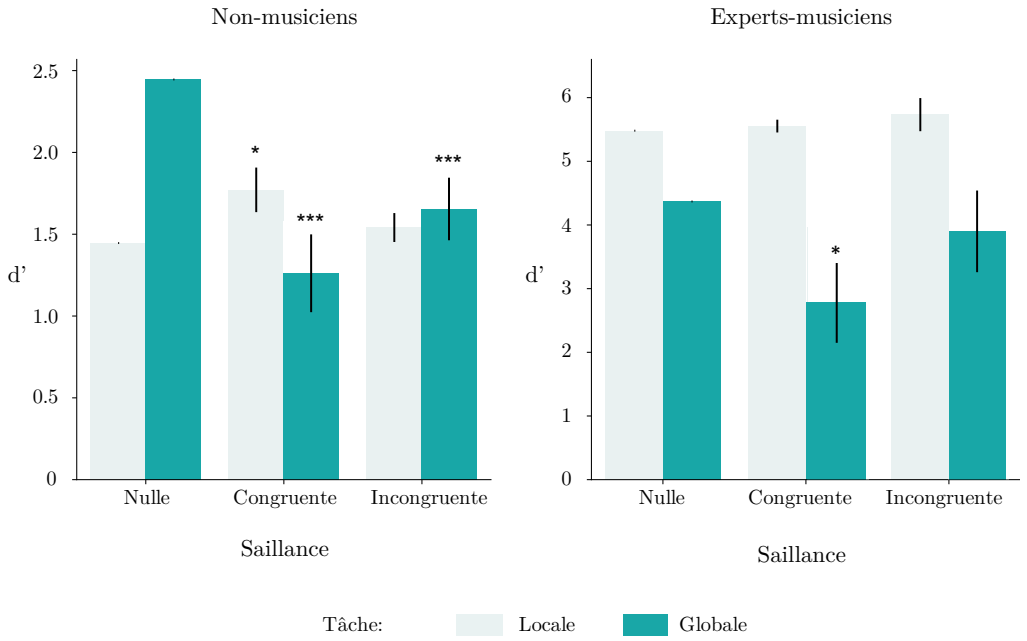

FIGURE 3.10 – Sensibilité (d') dans la tâche locale (bleu pastel) et la tâche globale (vert émeraude) pour les non-musiciens (à gauche) et les experts-musiciens (à droite) en fonction de la condition de saillance. Barres d'erreur : erreur standard de la distribution des scores des participants entre chaque condition et la condition sans saillance. La significativité des différences entre chaque condition et la condition sans saillance est indiquée par les étoiles (* : $p_{corr} < 0,05$, *** : $p_{corr} < 0,001$)

Pour les non-musiciens

Chez les non-musiciens, la présence de sons localement saillants implique une dégradation significative des performances dans la tâche globale, qu'elle soit congruente (T(10) = 4,99, p <0,001, p_{corr}<0,001, cohen-d = 1,51, power = 1,0) ou incongruente (T(10) = 4,65, p<0,001, p_{corr}<0,001, cohen-d = 1,4, power = 0,99).

Dans la condition de saillance congruente, on observe une amélioration significative de la performance dans la tâche locale (T(10) = 2,39, p = 0,019, p_{corr}=0,025, cohen-d = 0,72, power = 0,72). De fait, dans cette condition, leur sensibilité devient plus importante (d' = 1,77) dans la tâche locale que dans la tâche globale (d' = 1,26). Dans la condition incongruente, la performance dans la tâche locale n'est pas affectée significativement.

Pour les experts-musiciens

Chez les musiciens, la présence de sons localement saillants entraîne également une dégradation des performances dans la tâche globale, de manière significative dans la condition de saillance congruente (T(5) = 2,54, p = 0,026, p_{corr}=0,026, cohen-d = 1,04, power = 0,70). L'effet sur la tâche locale n'est pas significatif, de même que dans la condition de saillance incongruente. Ce résultat est à prendre avec précaution car il semblerait que la performance de certains experts ait pu plafonner par moments dans cette tâche en particulier. On observe par exemple en figure 3.8 que la sensibilité dans la tâche locale en condition de saillance nulle approche le maximum pour trois d'entre eux (le maximum est à 6,2).

On peut visualiser l'effet de la saillance sur l'avantage global dans le plan (GA, GL) en figure 3.11.

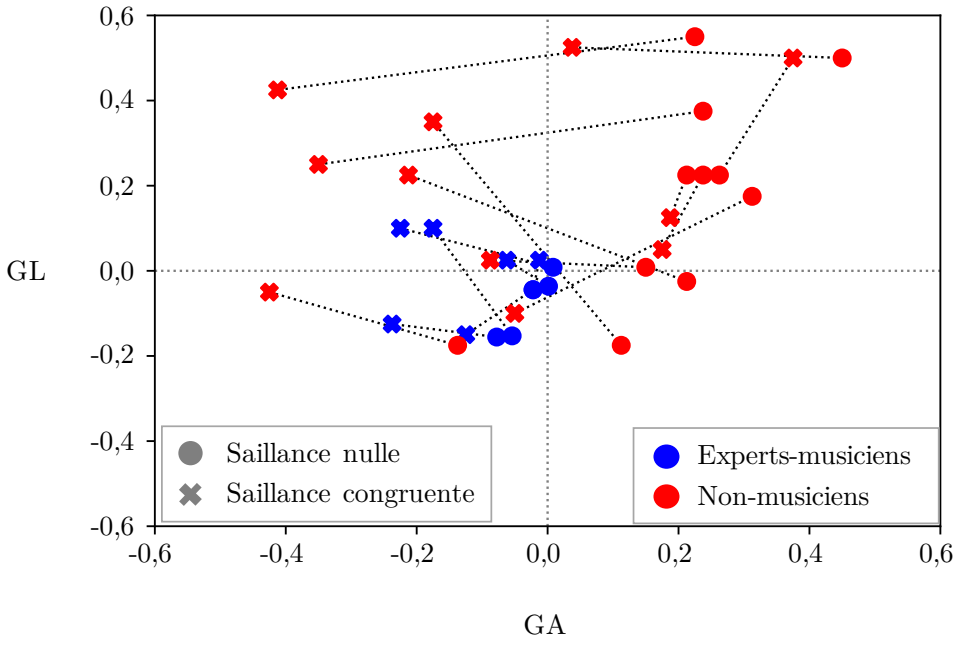

FIGURE 3.11 – Répartition des participants dans le plan (GA, GL), dans la condition de saillance nulle (ronds) et de saillance congruente (croix). Les musiciens sont en bleu, les non-musiciens en rouge.

On observe alors la répartition des participants en fonction de leur groupe d'appartenance dans la condition de saillance nulle et dans la condition de saillance congruente. Les déplacements des non-musiciens vers la gauche caractérisent l'inversion de l'avantage global observé en cas de saillance congruente avec la modification à détecter. Les experts-musiciens sont également concernés par un déplacement, dans une moindre mesure.

On peut observer les résultats dans le plan $(d'_{local}, d'_{global})$ en figure 3.12 pour observer la cause du décalage vers la gauche observé en figure 3.11, c'est-à-dire la diminution de l'avantage global. L'effet de la saillance s'y caractérise cette fois par un déplacement vers le bas pour la dégradation de la performance dans la tâche globale, et un déplacement vers la droite pour l'amélioration de performance dans la tâche locale pour les non-musiciens (l'effet de plafonnement pour les experts concernés y est également visible).

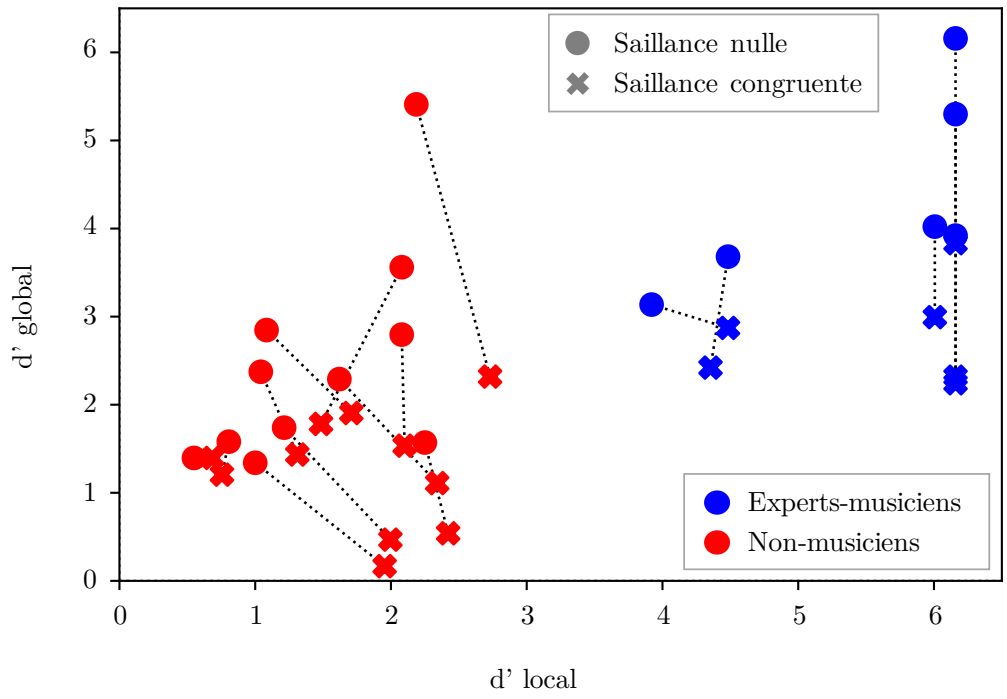

FIGURE 3.12 – Répartition des participants dans le plan $(d'_{local}, d'_{global})$, dans la condition de saillance nulle (ronds) et de saillance congruente (croix). Les musiciens sont en bleu, les non-musiciens en rouge.

3.3.3 Discussion

Effet de la saillance sur l'organisation du traitement local/global

Les performances des participants sur les essais contenant des manipulations de saillance sont affectées par rapport à celles obtenues sur les essais sans saillance. Cela confirme que l'effet de capture attentionnelle généré par une augmentation de brillance étudié au chapitre 2 est toujours significatif dans ce nouveau contexte. Ce dernier implique une analyse auditive plus poussée de la part des participants qui doivent ici traiter des stimuli plus complexes que précédemment. Les stimuli du paradigme du singleton additionnel mis en oeuvre au chapitre précédent ne contenaient que cinq sons de même hauteur diffusés à intervalle régulier. Ici, les participants perçoivent deux mélodies de 9 notes placées à intervalles plus variés ainsi que des variations de hauteur et de timbre, et doivent comparer des structures à différents niveaux. Malgré une montée en complexité de la scène sonore perçue et du traitement cognitif mis en oeuvre, on note donc que la saillance a toujours un effet sur les performances. Autrement dit, l'effet du mécanisme bottom-up qu'est la saillance, objet de nos travaux, perdure à plus haut niveau.

Plus précisément, on observe que les performances des non-musiciens dans la tâche globale sont toujours diminuées lorsqu'une information est localement saillante dans la mélodie. Que celle-ci soit congruente ou non avec la modification à détecter, sa présence dégrade l'appréciation des mélodies au niveau global. Ainsi, la saillance au niveau local attire l'attention sur ce niveau, toujours au détriment du niveau global.

De plus, lorsque cette saillance est congruente avec la modification locale à repérer, les participants sont plus performants dans le traitement de l'information locale. L'amélioration de la performance dans la tâche locale pour les non-musiciens lorsque la saillance est présente au niveau du triplet localement modifié montre que l'attention est bien orientée de manière privilégiée sur cette information au niveau local. La saillance n'est donc pas une simple "distraction" supplémentaire dans le stimulus, mais bien un guide attentionnel

qui permet d'en souligner une partie précise.

Il y a ainsi une réorganisation du traitement local/global de l'information sonore temporelle du fait de la présence de saillance locale. Ce résultat est cohérent avec celui observé pour le traitement des informations spatiales dans la modalité visuelle par Mevorach et al. (2006).

Interaction avec les mécanismes cognitifs sous-tendus par l'expertise musicale

Les résultats des experts-musiciens dans la conditions de saillance nulle confirment dans un premier temps que l'expertise musicale est associée avec la capacité à déployer des mécanismes cognitifs leur permettant d'orienter plus facilement que les non-musiciens leur attention sur le niveau de leur choix. Ce résultat est cohérent avec ceux déjà observés précédemment (Black et al., 2017 ; Ouimet et al., 2012 ; Susini et al., 2020).

Malgré cela, les experts-musiciens ne résistent pas complètement à l'attraction attentionnelle au niveau local causée par la saillance. En effet, la saillance locale renforce leur avantage local dans la condition congruente, principalement du fait d'une dégradation des performances dans la tâche globale. La capture attentionnelle au niveau local semble ainsi perturber le traitement au niveau global pour les musiciens comme pour les non-musiciens : les mécanismes cognitifs sous-tendus par l'expertise musicale ne suffisent pas à annuler cet effet.

L'effet de la saillance sur la performance dans la tâche locale n'est quant-à-lui pas significatif pour les experts-musiciens. Il ne l'est pas non plus dans la condition de saillance incongruente. Ces résultats semblent ainsi montrer que l'effet de la saillance locale est amoindri chez les experts-musiciens, mais des travaux plus poussés seraient nécessaires pour vérifier cette hypothèse. De fait, seuls 7 experts-musiciens ont participé à notre expérience. Une piste envisageable serait d'analyser la performance de participants en fonction de

leur niveau d'expertise, du non-musicien à l'expert en passant par le musicien amateur, comme chez Susini et al. (2023). Par ailleurs, il semblerait que, pour certains experts, les performances plafonnent par moment (notamment dans la tâche locale). Ainsi, si une dégradation de leurs performances dans la tâche globale a pu être observée, un paradigme impliquant une tâche plus complexe (notamment des profils mélodiques plus complexes) pourrait révéler un effet dans la tâche locale également, comme pour les non-musiciens.

L'organisation du traitement local/global est ainsi guidée par la saillance des stimuli. Autrement dit, l'analyse des scènes sonores diffusées dans ce paradigme est affectée par les caractéristiques des sons qu'elles contiennent. Cet effet est observé même en présence de mécanismes descendants visant à l'inhiber.

Ces résultats suggèrent que l'analyse d'une scène auditive est bien affectée par le mécanisme de capture attentionnelle mesuré au chapitre 2. Un phénomène comme celui de primauté du traitement holistique est en effet modulé par la présence de sons localement saillants. Dès lors, on peut se demander ce qu'il en est concernant la perception de scènes encore plus complexes. C'est l'objet du chapitre 4, dans lequel nous nous intéressons à l'effet de la présence d'évènements localement saillants sur la perception de l'agrément sonore dans des scènes sonores environnementales.

3.4 Conclusion

Nous avons ainsi montré que la capture attentionnelle affecte la hiérarchie du traitement de l'information sonore temporelle, en favorisant le traitement au niveau auquel se situe l'information saillante. Le phénomène de primauté du traitement holistique se trouve ainsi inversé lorsque des éléments sont saillants au niveau local. Cet effet de la saillance est également relevé en présence de mécanismes descendants sous-tendus par l'expertise musicale et visant à l'inhiber.

Nous avons ainsi confirmé qu'une des variations du timbre (la brillance) étudiée au chapitre 2 dans des séquences sonore simples induisait un effet de capture attentionnelle dans des séquences sonores plus complexes. Cela confirme l'importance du timbre comme déterminant de la saillance, et confirme que l'analyse de scènes auditives est affectée par la présence de sons saillants, même lorsqu'elle interagit avec d'autres mécanismes descendants.

Cette étude invite à poursuive l'exploration de l'effet de mécanismes ascendants sur l'analyse de scènes auditives. Entre autres, le paradigme ici mis en oeuvre pourrait être enrichi avec une tâche adaptative dont la complexité serait ajustée au niveau d'expertise de chaque participant. Une étude similaire mettant en jeu la perception de différents flux sonores simultanés, des mélodies entrelacées par exemple, pourrait également être une piste intéressante.

L'influence d'évènements saillants sur l'analyse et le traitement de scènes auditives ayant été mise en évidence, il pourrait être intéressant d'observer comment cet effet affecte la perception de scènes sonores plus complexes. Or, nous avons constaté au chapitre 1 que la perception et l'appréciation de paysages sonores mettait en jeu des mécanismes complexes et des concepts variés (cf. section 1.4). Nous avons notamment souligné que les indicateurs physiques visant à prédire la qualité d'un environnement sonore devaient être accompagnés d'une prise en compte du rôle des sources sonores le composant (cf. section 1.4.3). Il semble ainsi intéressant d'étudier le rôle de la saillance auditive dans la perception et l'appréciation de paysages sonores. C'est l'objet du chapitre 4.

Chapitre 4

Perception de scènes sonores environnementales : saillance et désagrément

"Le trop d'attention qu'on a pour le danger fait le plus souvent qu'on y tombe."

Jean de La Fontaine

4.1 Introduction

C'est autour de la problématique d'exposition au bruit que sont nées les motivations pour mener ces travaux de thèse (cf. Introduction). Plus précisément, nous nous demandions comment la saillance dans une scène sonore environnementale est susceptible d'en affecter la perception et l'appréciation. Le chapitre précédent nous a permis de montrer que la présence de sons saillants entraîne une réorganisation de l'analyse auditive de la scène qui les contient. L'effet de la saillance a été étudié pour des séquences sonores mélodiques contrôlées expérimentalement. Certes, l'analyse des séquences, complexes dans leur organisation temporelle, impliquait une comparaison de profils de hauteur variés et une hiérarchie du traitement de l'information entre niveau local et niveau global. Cependant, les stimuli étaient encore dépourvus de toute complexité sémantique et ne faisaient pas intervenir de flux auditifs différents et simultanés, comme c'est le cas dans les scènes environnementales [1].

Nous avons vu en partie 1.4 que des paramètres acoustiques ne sont que partiellement associés à la perception des paysages sonores, et que l'identification des sources sonores qui les composent est déterminante pour mieux les caractériser. La saillance, médiatrice de l'émergence des sources sonores dans leur contexte, peut-elle aider à mieux prédire l'appréciation des paysages sonores ?

Nous nous demandons ainsi dans ce chapitre comment l'appréciation de paysages sonores pourrait être affectée par la saillance auditive. Plus précisément, nous souhaitons observer comment la saillance affecte des évaluations d'agrément/désagrément [2] sonore dans des scènes sonores environnementales.

1. On parlera de "scènes environnementales" pour désigner des scènes naturelles complexes représentant des scénarios potentiels du monde réel (Huang et al., 2017)
2. On parlera d'évaluations d'agrément sonore ou de désagrément sonore de manière équivalente, puisque il s'agit d'une même mesure sur la dimension agréable/désagréable.

4.1.1 Évaluations de saillance

Toutes les évaluations de saillance réalisées dans ce chapitre ont été faites par le modèle prédictif introduit par Huang et al. (2017). Le choix de ce modèle a été fait en raison de sa conception à partir de mesures de saillance perceptives sur des scènes sonores environnementales. Ce modèle transforme en premier lieu le signal audio en spectrogramme, pour pouvoir en extraire des caractéristiques acoustiques et psychoacoustiques : entre autres sonie, hauteur, harmonicité, brillance, rugosité, etc. Toutes ces caractéristiques sont ensuite normalisées, puis recombinées pour en déduire une courbe temporelle de saillance.

Pour entraîner ce modèle, Huang et al. (2017) ont demandé à des participants, à qui l'on présentait une scène sonore dans chaque oreille (écoute dichotique), d'indiquer où leur attention était portée. Pour chaque scène, la réponse moyenne était obtenue en moyennant les réponses obtenues sur toutes les confrontations entre cette scène et une autre scène du corpus pour tous les participants. Les pics dans la dérivée de cette réponse moyenne étaient considérés comme correspondant aux évènements saillants. Puis, chaque scène était découpée en segments d'une seconde avec un chevauchement de 0,75 secondes. Pour chaque caractéristique, si un pic était observé dans sa dérivée, le segment contenant le pic était annoté. Enfin, une analyse discriminante linéaire a permis de combiner les prédictions de chaque caractéristique, sur chaque segment, en une prédiction globale visant à s'approcher le plus possible des données comportementales. Les poids de cette analyse linéaire correspondent à la contribution de chaque caractéristique dans la prédiction de la saillance.

Ainsi, ce modèle prend en entrée le signal audio d'une séquence sonore et renvoie une courbe que l'on peut interpréter comme l'évaluation de la saillance de cette séquence en fonction du temps. Les données issues du modèle sont normalisées par scène analysée. De fait, on ne peut comparer la saillance de manière absolue entre des sources issues de scènes différentes.

Cette contrainte est d'ailleurs fondée : une source sonore est toujours incluse dans un certain contexte, et on ne pourrait évaluer sa saillance de manière absolue : on ne peut comparer la saillance de sources que si elles font partie de la même scène.

Le modèle de saillance ayant été entraîné sur des paysages sonores variés (Kothinti et al., 2021), il est raisonnable de l'appliquer sur les paysages sonores que nous avons utilisés dans cette étude.

4.1.2 Mesure de désagrément continu

Si nous pouvons ainsi disposer de prédictions de saillance au cours du temps pour des scènes sonores environnementales, il faut pouvoir évaluer l'agrément sonore perçu en fonction du temps sur ces mêmes scènes afin d'étudier le lien potentiel entre les deux. Cette volonté d'évaluer cette grandeur en continu, et de ne pas se limiter à une mesure globale pour chaque scène (qui a été néanmoins été évaluée et rapportée en annexe 6), est motivée par plusieurs facteurs :

— Une même scène peut contenir différents évènements sonores saillants : seule une évaluation continue permet d'observer l'effet de chaque évènement sur le désagrément. Un jugement global exigerait d'aggréger toutes les informations issues de la courbe de saillance (nombre de pics, hauteur des pics, etc.) pour essayer de les lier à la note globale de désagrément.
— Une part du lien entre saillance et désagrément pourrait se trouver entre leurs dérivées temporelles (les variations temporelles de l'une affecterait les variations temporelles de l'autre) : seule une mesure en continu permettrait d'identifier ce lien
— Une restitution globale rétrospective est affectée par différents effets, notamment un effet de récence - le début et la fin des séquences sont mieux mémorisés que le reste (Västfjäll, 2004) - et un effet d'"apogée" - le jugement global est influencé par le moment le plus intense de la

scène (Fiebig and Sottek, 2015). Un jugement global ne donnerait donc pas exactement une estimation de l'agrément moyen de la séquence, mais plutôt une évaluation du moment le plus intense ainsi que du début et de la fin.

— Une évaluation en continu est plus proche de l'expérience des usagers immergés dans un paysage sonore : ce sont les sources saillantes qui attirent, par moments, l'attention de l'auditeur sur son environnement sonore et lui permettent de l'évaluer (cf. 1.4.4). Il est donc souhaitable d'étudier leur effet au moment où elles surviennent.

Ainsi, nous formerons un corpus de paysages sonores variés, dont nous obtiendrons d'une part des évaluations de saillance via un modèle prédictif, et d'autre part des mesures de désagrément via une expérience d'évaluation en continu. Nous mènerons ensuite des analyses pour tenter de comprendre comment la saillance affecte les évaluations de désagrément sonore dans ces scènes environnementales.

4.1.3 Lien entre saillance et désagrément

Une étude s'intéressant au lien entre saillance temporelle et évaluation continue de désagrément a été menée par Filipan et al. (2019). Aumond et al. (2017b) avaient d'abord obtenu une évaluation de désagrément par des participants en leur présentant des montages sonores issus d'enregistrements de balades sonores en ville. Ces travaux leur avaient permis d'observer un effet de récence dans les évaluations, ainsi qu'un lien entre niveau sonore et désagrément instantanés. Filipan et al. (2019) ont alors repris ces données de désagrément pour les lier à une mesure de saillance. Plus précisément, ils ont créé leur propre mesure, basée sur les modulations spectro-temporelles et développée spécifiquement pour cette étude, et observé comment elle était liée à la probabilité d'observer un changement dans les évaluations de désagrément par les participants. Les résultats ont mis en évidence que leur mesure de saillance prédisait mieux que le niveau sonore les changements d'évaluation de désagrément dans leur corpus. Leur méthodologie comportait

cependant quelques limites :

— Leur corpus se concentrait sur un environnement sonore urbain en particulier (issu de balades dans Paris) et plus précisément sur l'effet de transitions dues au passage entre différentes zones urbaines : zones calmes (parc, rue piétonne) et zones bruyantes (boulevards, rues passantes). Ces transitions introduisaient ainsi de grandes variations, de niveau sonore notamment, qui déterminaient une grande part de la variabilité des grandeurs étudiées : niveau, saillance et désagrément. Face à cette grande variabilité, la contribution des sources sonores propres à l'environnement était amoindrie. Un exemple est présenté en figure 4.1. On observe la transition, vers la 120^{e} seconde, entre une rue piétonne et une rue passante. Cette transition induit des variations d'agrément et de niveau sonore bien supérieures à celles de passages de véhicules (un deux-roues à la 160^{e} seconde par exemple), pourtant clairement saillant à l'écoute (et d'après le modèle de saillance de Huang et al. (2017) appliqué à cette séquence). Leurs résultats étaient ainsi empreints de cet effet de transition entre les différentes zones, et l'effet des sources sonores dans leur environnement était potentiellement amoindri.

— Par ailleurs, leur modèle de saillance n'a pas été validé par des données de saillance issues de mesures perceptives recueillies auprès de participants. Bien que prédisant mieux les changements d'évaluations de désagrément que le niveau sonore dans leur étude, cette mesure n'est donc pas validée du point de vue de la saillance auditive.

Nous proposons dans ces travaux de thèse d'étudier l'effet de la saillance sur l'appréciation de paysages sonores plus variés que les environnements urbains, et en ne considérant pas de changement d'environnement au sein d'une même séquence sonore, pour observer l'influence des sources faisant partie intégrante du paysage. De plus, nous souhaitons utiliser un modèle de saillance ayant été conçu spécifiquement à partir de données perceptives.

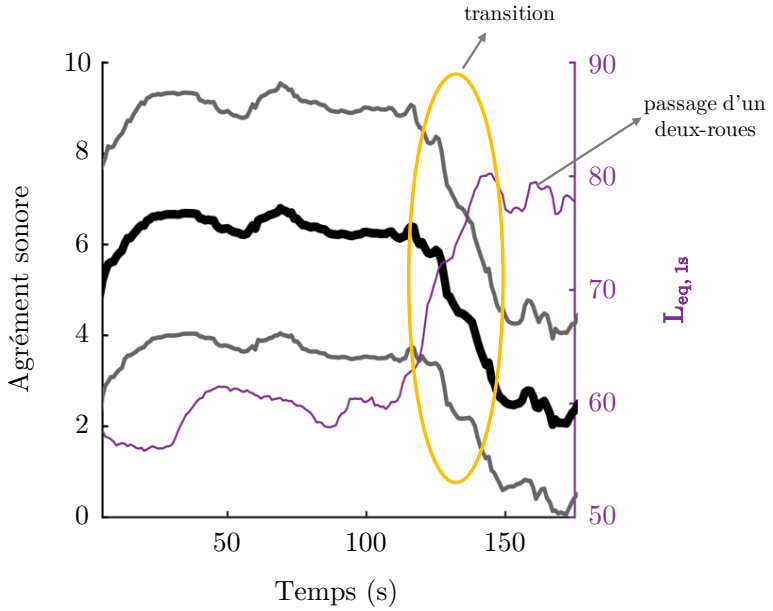

FIGURE 4.1 – Exemples de scène sonore (S2) diffusée et évaluée par Aumond et al. (2017b), puis reprise par Filipan et al. (2019). La ligne noire épaisse représente l'évaluation moyenne d'agrément, les lignes noires plus fines les écarts-types associés, et la courbe violette représente le niveau sonore ($L_{eq,1s}$).

Pour ce faire, nous proposons de mettre en oeuvre une expérience de mesure de désagrément en continu, similaire à celle utilisée par Aumond et al. (2017b), puis étudier sa relation avec la saillance selon une méthodologie similaire à celle de Filipan et al. (2019) en utilisant les prédictions de saillance du modèle de Huang et al. (2017).

4.2 Expérience

4.2.1 Participants

36 participants (dont 17 hommes) âgés entre 18 et 35 ans (moyenne de 25 ans) ont volontairement pris part à l'expérience. Aucun n'a déclaré avoir des problèmes d'audition. Ils ont donné leur consentement par écrit avant l'expérience et ont été rémunérés pour leur participation.

4.2.2 Équipement

Les stimuli ont été présentés aux auditeurs par l'intermédiaire d'un casque Beyerdynamic DT-770 PRO. L'expérience s'est déroulée dans les cabines insonorisées de l'*INSEAD-Sorbonne Université Behavioural Lab*. L'interface de test a été codée avec le logiciel Max (v8) [3] et s'est déroulée sur des Mac Mini. Les participants évaluaient l'agrément sonore en utilisant un contrôleur midi conçu pour cette expérience. Plus précisément, un curseur était relié à une carte arduino dans un boîtier connecté par USB à l'ordinateur (voir annexe 4 pour le détail sur les boîtiers). L'interface est présentée en figure 4.2.

4.2.3 Stimuli

Comme nous souhaitons étudier l'effet de sources saillantes sur la perception du paysage sonore, il faut s'assurer que les scènes de notre corpus contiennent bien des évènements saillants. Nous devons donc construire un corpus en suivant une méthodologie fondée sur les observations des prédictions de saillance. Pour assurer une diversité des effets observés, la construction du corpus doit également prendre en compte les résultats de la littérature concernant les effets supposés positifs ou négatifs de certaines sources sur l'agrément sonore (Guastavino, 2006 ; Lavandier and Defréville, 2006 ; Nilsson and Berglund, 2006).

3. https://cycling74.com/products/max

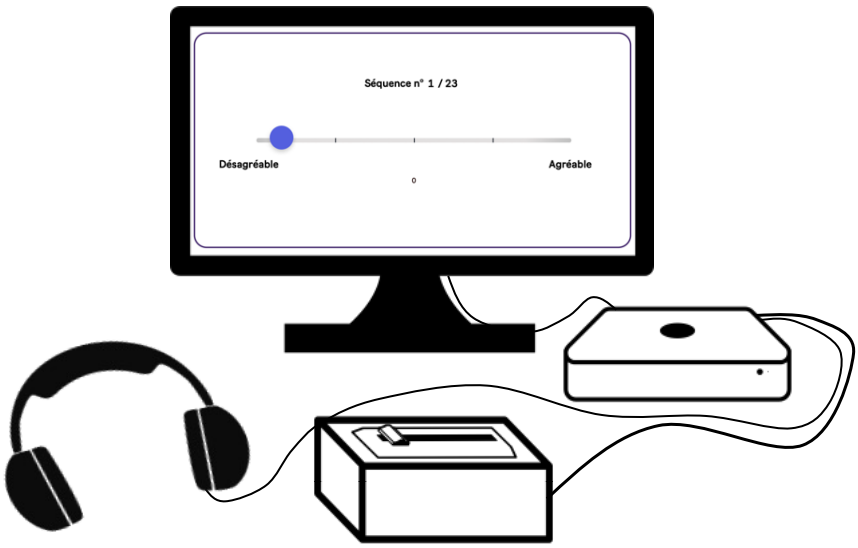

FIGURE 4.2 – Équipement et interface : le curseur à l'écran, présenté dans l'environnement Max MSP, est contrôlé par les déplacements du curseur sur le boîtier.

Le corpus de stimuli diffusés regroupe 23 scènes sonores, issues de 11 environnements différents présentés en table 4.1. Ces environnements sont issus de la base de données sonore Sound Ideas [4].

Construction des stimuli

Chaque environnement sonore a donné lieu à 2 scènes (ou 3 pour l'un d'entre eux) de 56 secondes : une contenant un événement saillant au début de l'extrait, l'autre non. L'identification des événements saillants dans les enregistrements a été déterminée en utilisant le modèle de saillance auditive de Huang et al. (2017).

Toutes les scènes suivent une construction commune (voir figure 4.3 pour des exemples) : une "introduction" d'une dizaine de secondes, un passage de quelques secondes que nous appelons "passage-clé" à partir de la dixième seconde environ, puis le reste du déroulement de la scène. Les scènes d'une même paire sont identiques à l'exception de ce passage clé, qui contient soit un événement saillant soit rien de particulier. Aucune manipulation dans le

4. https://www.sound-ideas.com

contenu des enregistrements n'a été effectuée, à l'exception d'un cross-fade pour introduire le passage clé. Pour chaque paire, l'introduction, le déroulement de la scène et le passage clé sont issus du même environnement sonore.

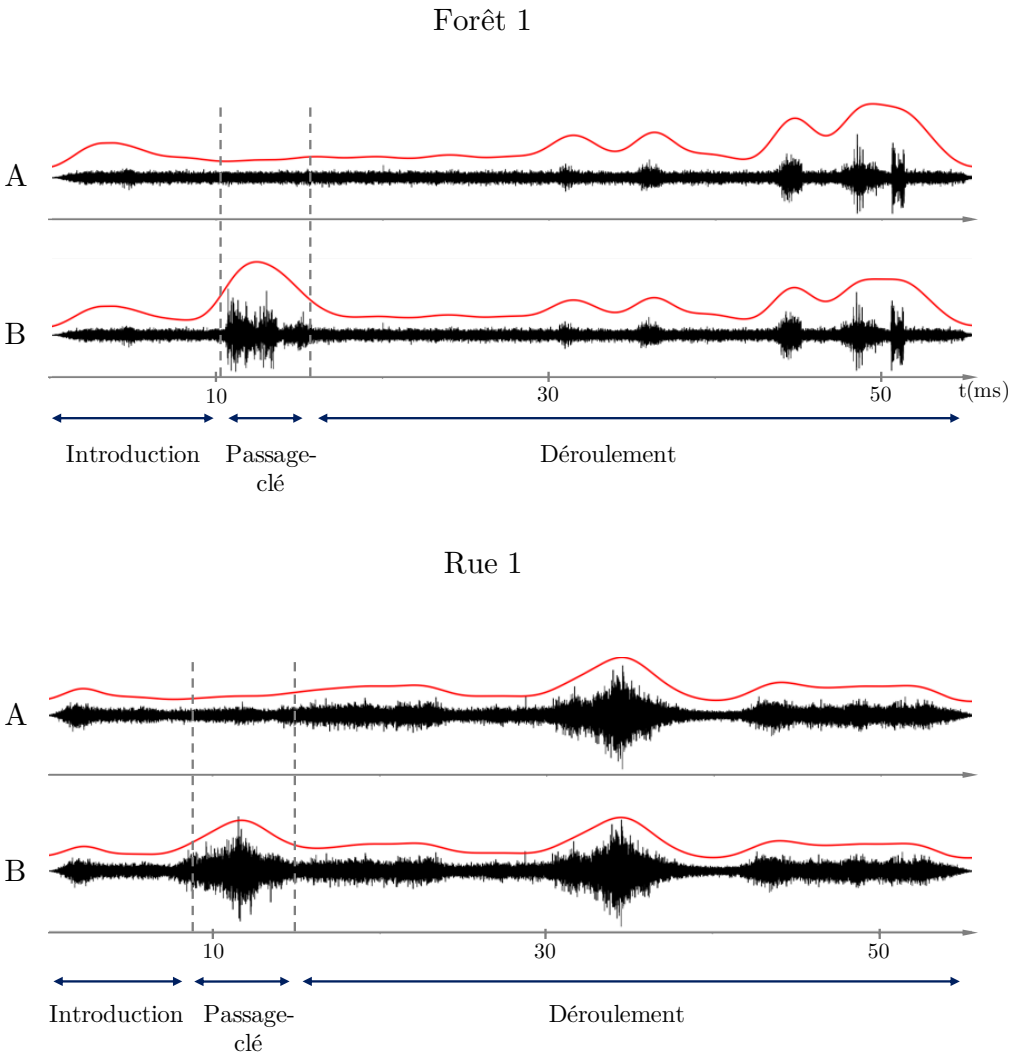

FIGURE 4.3 – Exemples de scènes sonores diffusées dans l'expérience, pour deux environnements différents : la forêt 1 et la rue 1. En noir, la forme d'onde, en rouge, la courbe de saillance issue du modèle de Huang et al. (2017). Pour chaque environnement, la scène A est celle ne contenant pas d'événement saillant dans le passage-clé, la scène B en contient un : un cri d'écureuil pour la scène B en forêt, le passage d'un deux-roues pour la scène B dans la rue.

L'introduction dure une dizaine de secondes et ne contient pas d'évènement saillant. Sa durée a été prévue pour laisser le temps aux participants de percevoir l'ambiance de la scène au début de la tâche d'évaluation continue de l'agrément sonore, et d'ajuster la position du curseur avant la survenue d'un évènement saillant potentiel. Le caractère amorphe de l'introduction assure que les participants ne sont pas immédiatement affectés par la saillance de sources sonores pendant le repositionnement du curseur initial.
Dans la scène contenant un événement saillant, c'est au moment du passage-clé que ce dernier se produit, en moyenne à 10,9 secondes (écart-type de 1 seconde) du début. Dans la scène n'en contenant pas, l'ambiance amorphe de l'introduction est poursuivie sur la durée du passage-clé.
Le reste du déroulement de la scène n'a pas été contrôlé. Sur une durée variable suivant le passage-clé (25 secondes en moyenne, écart-type de 4 secondes), il n'y a pas d'autre évènement aussi saillant que l'évènement saillant du passage-clé selon le modèle de Huang et al. (2017), puis d'autres évènements peuvent survenir.

Sélection des stimuli

Des environnements variés ont été choisis pour concevoir les scènes : chantier de travaux, rues, port ou parc pour des environnements urbains ; campagne, forêt, désert pour des environnements plus naturels. Les évènements sonores saillants présents dans les passages-clés des scènes associées étaient des évènements présents dans les environnements (puisqu'aucune manipulation n'a été faite sur le contenu des environnements sonores). La répartition assurait la présence dans le corpus d'environnements et de sources sonores à la fois agréables et désagréables a priori. Cette hypothèse a été vérifiée lors de l'analyse des résultats. Les niveaux de restitution respectaient les écarts relatifs des enregistrements originaux. Les niveaux équivalents pondérés A des scènes totales étaient ainsi distribués sur la plage $L_{A,\,eq,\,scène} \in [55\,;\,67]$ dB(A) (moyenne : 61 dB(A), écart-type : 4 dB(A)). Les différents environnements sonores et les évènements saillants associés sont listés en table 4.1.

Scène	Événement dans le passage clé
Chantier ‘amorphe’	-
Chantier	Scie électrique
Rue 1 ‘amorphe’	-
Rue 1	Passage d’un deux-roues
Rue 2 ‘amorphe’	-
Rue 2	Klaxon de voiture
Rue 3 ‘amorphe’	-
Rue 3	Chuintement d’un véhicule
Parc ‘amorphe’	-
Parc	Chant d’un oiseau
Port ‘amorphe’	-
Port 1	Corne de brume
Port 2	Chant de mouettes
Campagne ‘amorphe’	-
Campagne	Chant d’oiseau
Désert ‘amorphe’	-
Désert	Chant d’un hibou
Forêt 1 ‘amorphe’	-
Forêt 1	Cri d’un écureuil
Forêt 2 ‘amorphe’	-
Forêt 2	Arrêt du chant d’un grillon
Salle de réception ‘amorphe’	-
Salle de réception	Tintement de verres

TABLE 4.1 – Liste des scènes sonores du corpus. Les scènes notées ‘amorphe’ sont celles ne contenant pas l’évènement saillant dans le passage-clé. Elles sont associées aux scènes de la même case qui, elles, contiennent un évènement saillant dans le passage-clé.

4.2.4 Procédure

Les scènes étaient présentées dans un ordre aléatoire pour chaque participant. La consigne était la suivante :

Au cours de cette expérience, vous écouterez 23 scènes sonores variées, d'une durée de 50 secondes chacune environ. Vous avez devant vous un curseur, positionné sur un boîtier allant de « désagréable » (à gauche) à « agréable » (à droite). Ce boîtier est relié à l'ordinateur et les déplacements de curseur sur le boîtier seront retranscrits sur l'écran.
Des scènes sonores vont être diffusées dans le casque. Vous devrez ajuster continuellement la position du curseur sur le boîtier pour évaluer l'agrément sonore de la séquence entendue : à quel point il vous semble agréable/désagréable. Plus l'environnement sonore vous sera agréable, plus vous déplacerez le curseur vers la droite ; plus il sera désagréable, plus vous déplacerez le curseur vers la gauche.

À la fin de chaque scène, le participant devait déclencher le lancement de la scène suivante. L'expérience durait 30 minutes.

4.3 Analyse et résultats

4.3.1 Données

4.3.1.1 Évaluations de désagrément

Les données d'évaluation de l'agrément sonore ont été collectées à une fréquence de 60 Hz. Les données issues du modèle de saillance avaient une fréquence d'échantillonnage de 15,6 Hz. Toutes les données d'évaluation d'agrément ont donc été sous-échantillonnées, à la fréquence de celles des données de saillance.

Les données de désagrément consistent donc en un corpus de 828 séries temporelles correspondant aux déplacements de curseur pour chaque scène et chaque participant. Les données midi issues du boîtier, comprises entre 0 et 127, ont subi la transformation $x \mapsto 10 * (1 - \frac{x}{127})$). Les valeurs sont donc comprises entre 0 et 10, 0 correspondant à l'évaluation la plus agréable, 10 à la plus désagréable. Certains participants n'ayant pas bougé le curseur sur toute la durée de certaines scènes, les fichiers correspondants ont été retirés : 8,7% des données ont été retirées à ce stade. Puis, pour chaque scène sonore, les réponses des participants ont été moyennées. Des exemples de courbes de désagrément moyennées sont présentés en figure 4.4. Les scènes contenant l'évènement saillant dans le passage-clé sont présentées ici pour chaque environnement.

Le curseur étant manipulé physiquement par les participants, sa position au début de chaque scène sonore correspondait à la position où il se situait à la fin de la scène évaluée précédemment. Ainsi, un temps d'intégration était nécessaire au début de chaque scène pour replacer le curseur à une position qui convenait au participant pour cette scène. La distribution des scènes étant aléatoire pour tous les participants, la valeur observée au début de chaque scène correspond donc à la moyenne des évaluations de désagrément de la scène précédente par les 36 participants. Cela correspond à la moyenne de la dernière évaluation de 36 scènes piochées aléatoirement parmi les 23 du

corpus. Cette valeur, proche de 5 pour chaque scène, confirme que le corpus est assez équilibré entre scènes jugées agréables et scènes jugées désagréables.

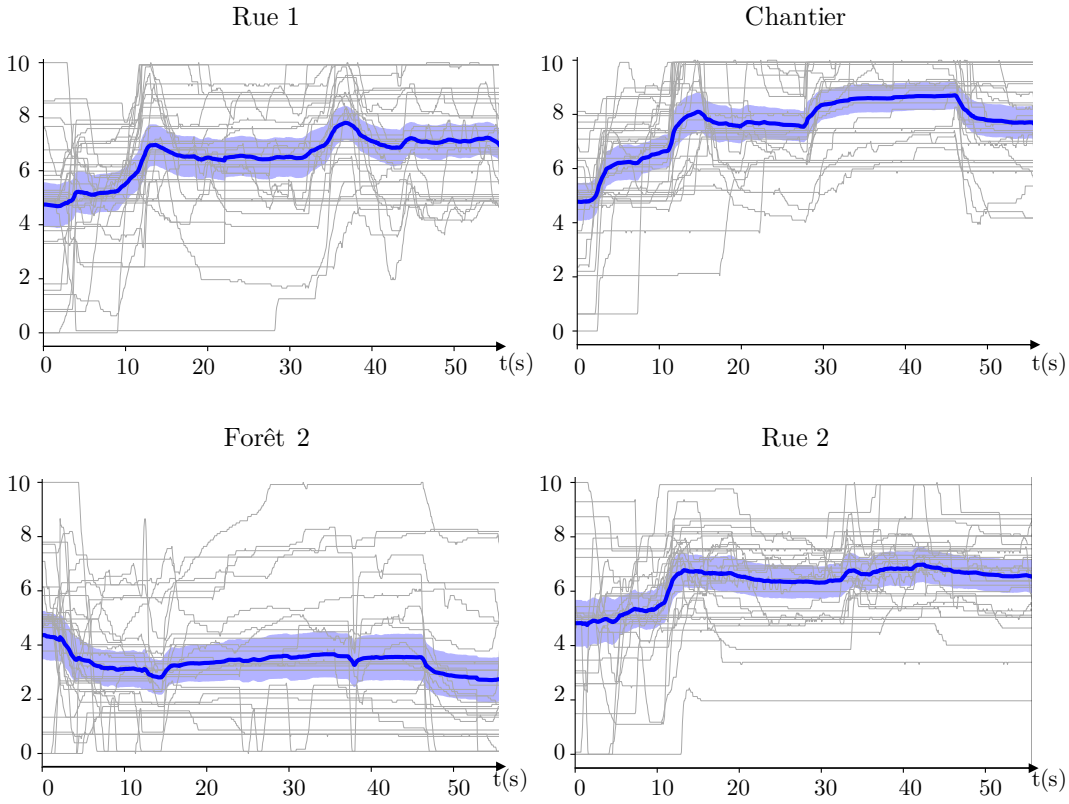

FIGURE 4.4 – Exemples de courbes de désagréments moyennées (en bleu) en fonction du temps, issues des évaluations de tous les participants (en gris) pour 4 scènes différentes, chacune étant la scène contenant un évènement saillant dans l'environnement précisé. La bande en bleue autour de la moyenne représente l'intervalle de confiance à 95%.

4.3.1.2 Données de saillance

Le modèle de saillance extrait les caractéristiques temporelles du signal pour en déduire la courbe de saillance. Nous avions accès à ces caractéristiques en plus de la courbe de saillance finale (cf. 4.1.1). Nous nous sommes particulièrement intéressés à la saillance (par définition), la sonie (indicateur évaluant le niveau sonore perçu, classiquement utilisé dans l'évaluation du

paysage sonore), la brillance et la rugosité (les deux attributs du timbre étudiés au chapitre 2). Ces mesures, extraites du modèle de Huang et al. (2017), sont les suivantes :

— saillance : la série temporelle en sortie du modèle ;
— sonie : la moyenne des enveloppes calculées sur 28 bandes de fréquences en Bark sur la plage [250 ; 12000] Hz ;
— brillance : le centre de gravité spectral ;
— rugosité : l'énergie moyenne aux vitesses de modulations appartenant à la plage [20 ; 100] Hz ;

Toutes les séries temporelles ont été lissées sur une fenêtre de 1,5 secondes, conformément à Huang et al. (2017), puis normalisées par scène (car la saillance dépend du contexte qui est propre à chacune d'entre elles, voir 4.1.1). Seuls les passages-clés des scènes sonores appairées peuvent être comparés en terme de saillance, car le contexte est exactement le même au moment du début du passage-clé (même introduction pendant les 10 premières secondes) pour ces scènes. On peut observer quelques exemples de ces séries temporelles en figure 4.5. Toutes ces courbes étant normalisées, les valeurs sur l'axe des ordonnées ne sont pas signifiantes.

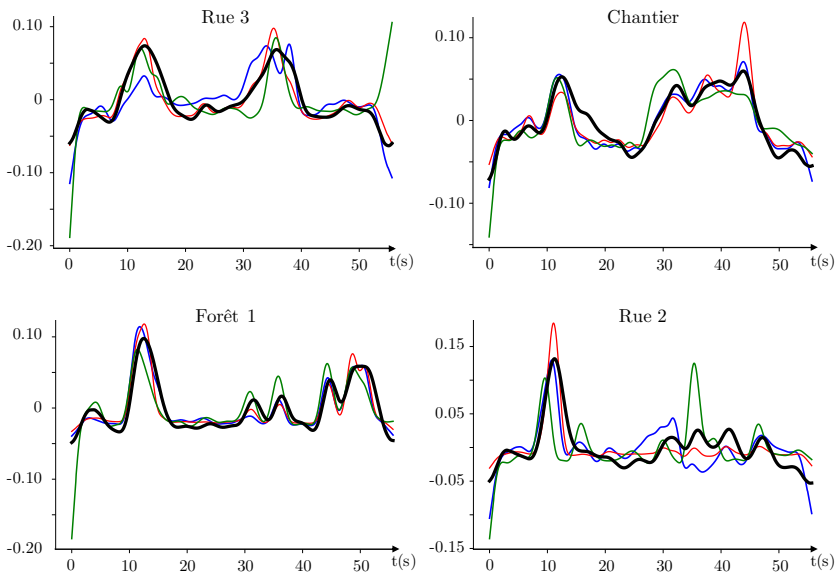

FIGURE 4.5 – Exemples de courbes de sonie (bleu), brillance (vert), rugosité (rouge) et saillance (noir) issues du modèle de Huang et al. (2017). Les séries sont normalisées par scène sonore.

4.3.2 Relations entre désagrément et saillance

Nous disposons donc des courbes de saillance renvoyées par le modèle, et des évaluations de désagrément moyennées en fonction du temps. Nous pouvons donc observer dans un premier temps, pour chaque scène sonore, la relation entre de ces deux grandeurs temporelles. Comme mentionné en section 4.3.1, les premières secondes étant dédiées au replacement du curseur, il a fallu ôter la partie de la scène correspondante. Nous avons donc mené les analyses sur les séries à partir de la 8^{e} seconde. L'évènement saillant survenant le plus tôt dans tout le corpus n'avait lieu qu'à la dixième seconde. Des exemples d'évolution de désagrément moyen et de saillance sont présentés en figure 4.6.

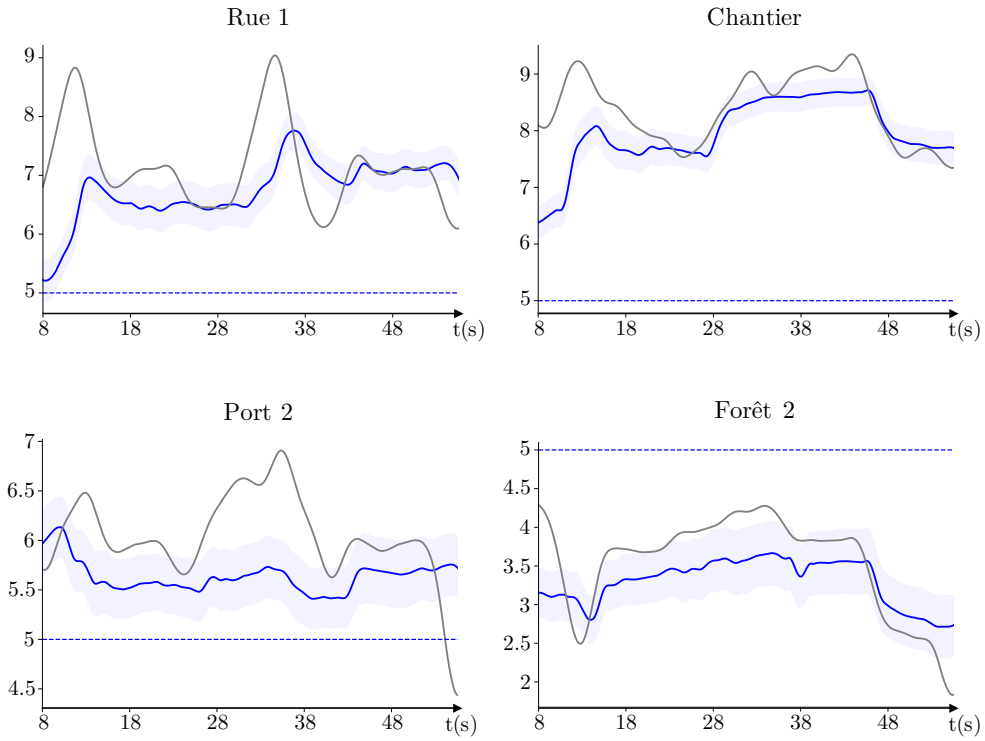

FIGURE 4.6 – Exemples de courbes de désagrément (bleu) et de saillance (gris). La bande en bleue autour de la moyenne représente l'erreur standard. L'ordonnée correspond aux valeurs de désagrément. Les valeurs de saillance étant relatives, seules leurs variations nous intéressent ici. Le trait en pointillés bleu indique la valeur 5, correspondant au milieu du curseur et séparant la zone d'évaluation désagréable dans les valeurs supérieures de la zone agréable dans les valeurs inférieures.

Visuellement, on constate une certaine correspondance entre l'évolution de la saillance et celle du désagrément. De fait, on observe de fortes corrélations entre les variations des deux séries temporelles. Des tests de Pearson entre la série de désagrément moyen et la série de saillance se sont d'ailleurs révélés significatifs pour 20 scènes sur les 23 du corpus (cf. annexe 5). Cependant, ce type d'analyse n'est pas assez spécifique des phénomènes observés et nous devons aller au-delà de simples observations de corrélations entre les séries. En effet, on observe ici que des variations de saillance semblent provoquer des variations de désagrément avec un certain délai. Les analyses menées plus loin (voir section 4.3.3) vérifieront ces observations en s'attachant aux relations de causalité pouvant exister entre les séries, en prenant en compte les délais pouvant exister entre leurs variations.

Écarts de désagrément et effet des évènements saillants

Le corpus de scènes sonores ayant été pensé en construisant les stimuli par paires (cf. scènes A et B en figure 4.3), nous pouvons nous intéresser à la comparaison des courbes de saillance et de désagrément des scènes appairées. Comme précisé en section 4.3.1.2, les seuls passages pouvant être comparés en terme de saillance entre deux scènes différentes sont les passages-clés des scènes appairées. De fait, ils surviennent après une introduction parfaitement identique. Un exemple est présenté en figure 4.7 pour les deux scènes issues de l'environnement "rue 1".

On constate sur cette figure 4.7 que les courbes de saillance reflètent bien la différence entre les deux scènes du même environnement : le passage clé autour de la dixième seconde présente un maximum local pour la scène contenant l'évènement saillant (en rouge), qui est absent dans l'autre scène (en bleu). Plus précisément, le premier pic de saillance en rouge est dû au passage d'un deux-roues, le suivant, en rouge ou bleu, au passage d'un camion.

Par ailleurs, on note avec intérêt que la suite des scènes, identique en tout point, donne lieu à deux profils similaires de saillance. Il convient cependant

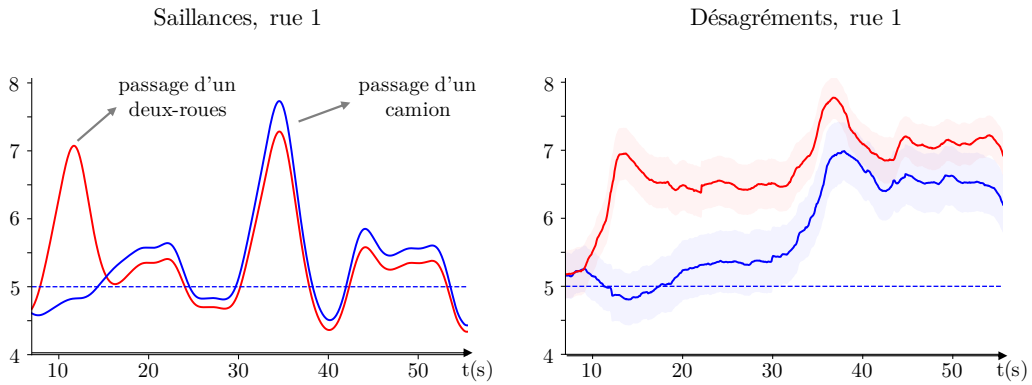

FIGURE 4.7 – Exemples de courbes de saillance (gauche) et des courbes de désagrément correspondantes (droite) pour les deux scènes de l'environnement rue 1. En bleu, la scène ne contenant pas d'évènement saillant dans le passage clé, en rouge, la scène en contenant un (le passage d'un deux-roues autour de la 10^{e} seconde). Les bandes en couleur bleue et rouge autour des courbes de désagréments représentent l'erreur standard sur la distribution des participants. L'ordonnée correspond aux valeurs de désagrément. Les valeurs de saillance étant relatives, seules leurs variations nous intéressent ici. Le trait en pointillés bleu indique la valeur 5, correspondant au milieu du curseur et séparant la zone d'évaluation désagréable dans les valeurs supérieures de la zone agréable dans les valeurs inférieures.

de noter qu'une légère différence persiste entre les deux : dans la scène ne contenant pas le passage du deux-roues à la dixième seconde, toute la suite du déroulement est prédite comme plus saillante. Ceci s'explique mathématiquement et perceptivement : mathématiquement, la courbe de saillance étant normalisée, les valeurs de la courbe bleue étant en moyenne inférieures dans la première partie de la scène, elles deviennent supérieures dans la deuxième partie après normalisation. Perceptivement, le contexte passé n'est plus le même dans les deux scènes après le passage-clé : les passages de véhicules dans la scène ne contenant pas le passage du deux-roues au début sont plus susceptibles d'être saillants puisqu'avant eux, rien n'a été particulièrement saillant.

Concernant les évaluations de désagrément, on observe un effet de l'évènement saillant : le passage du deux-roues provoque une augmentation de désagrément sur la courbe rouge et pas la bleue. L'évaluation du désagrément

dans la scène contenant cet évènement saillant reste supérieure à celle n'en contenant pas sur une durée supérieure à celle de l'évènement lui-même. Ce n'est qu'après un laps de temps suffisamment long, et notamment l'occurrence d'un deuxième évènement saillant (le passage d'un camion), que l'écart entre les deux courbes diminue. On peut observer la différence de désagrément induite par les différences de saillance entre les deux scènes en figure 4.8, sur laquelle nous ajoutons les mêmes courbes pour trois autres environnements.

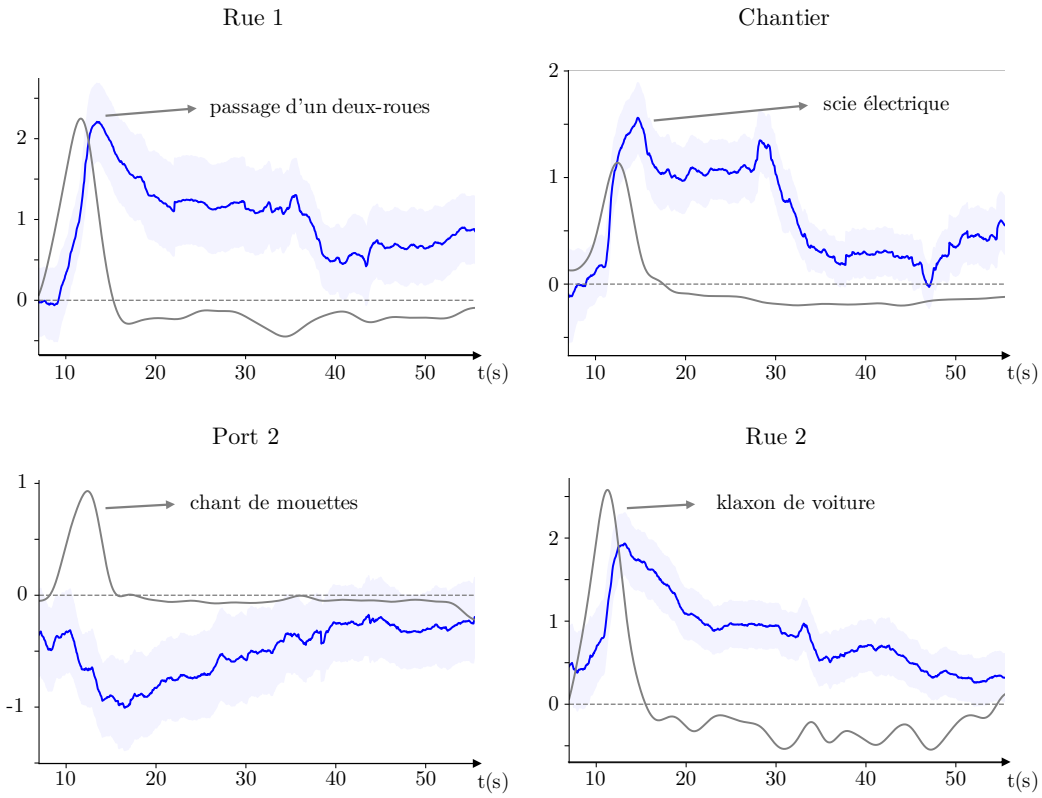

FIGURE 4.8 – Exemples de différences de désagrément (bleu) et de différences de saillance (en gris) pour les paires de scènes de quatre environnements. La bande en bleue autour des courbes de différences désagréments représente l'erreur standard sur la distribution des participants. L'ordonnée correspond aux valeurs de désagrément. Les valeurs de saillance étant relatives, seules leurs variations nous intéressent ici. Le trait en pointillés bleu indique la valeur 0 (pas de différence entre les deux scènes). Les évènements saillants sont le passage d'un deux-roues pour la rue 1, le bruit d'une scie électrique pour le chantier, le chant de mouettes pour le port 2, et un bruit de klaxon pour la rue 2.

Ces graphes (4.8) permettent de visualiser l'effet de chaque évènement saillant dans son environnement sonore sur l'évaluation de désagrément. Ainsi, on observe par exemple qu'un bruit de klaxon ou le passage d'un deux-roues dans la rue, ainsi que le bruit d'une scie électrique sur un chantier causent une hausse de désagrément perçu. Le chant des mouettes sur le port réduit au contraire ce désagrément.

Pour chaque environnement, on peut observer l'effet des différentes sources sur l'évaluation du désagrément. Pour ce faire, on calcule la distribution sur tous les participants de l'écart de désagrément entre les deux scènes de chaque environnement, moyenné sur t secondes, en commençant avec un délai de d secondes après l'évènement saillant. On peut alors en déduire une valeur moyenne de différence de désagrément engendrée par l'évènement saillant, que l'on peut comparer à 0 avec un test-t (la normalité des distributions est vérifiée à chaque fois, et une correction de Benjamini-Hochberg - dite du "False Discovery Rate" - est appliquée sur les résultats des multiples tests). Ces résultats sont présentés en figure 4.9, pour $t = 2$ secondes et $d = 1{,}5$ secondes. Ces valeur de t et de d montrent que 6 sources saillantes ont significativement affecté les évaluations de désagrément. Ces 6 sources ont un effet significatif dans un espace $d \in [0{,}5; 2{,}5]$ secondes et $t \in [-1{,}5 * d + 4{,}75; -1{,}5 * d + 5{,}5]$ secondes.

On peut classer ces sources en fonction de la direction de l'effet engendré sur l'évaluation de désagrément : si le désagrément diminue significativement en présence de la source, elle est caractérisée comme agréable, s'il augmente significativement, elle est caractérisée comme désagréable.

Dans les sources agréables, on retient :

- le chant des mouettes sur le port (T(27) = -2,29, p = 0,030, p_{corr} = 0,036, cohen-d = 0,43) ;
- l'arrêt du chant du grillon dans la forêt (T(27) = -2,19, p = 0,037, p_{corr} = 0,037, cohen-d = 0,41) ;
- le chant d'oiseau à la campagne (T(25) = -2,53, p = 0,017, p_{corr} = 0,026, cohen-d = 0,50) ;

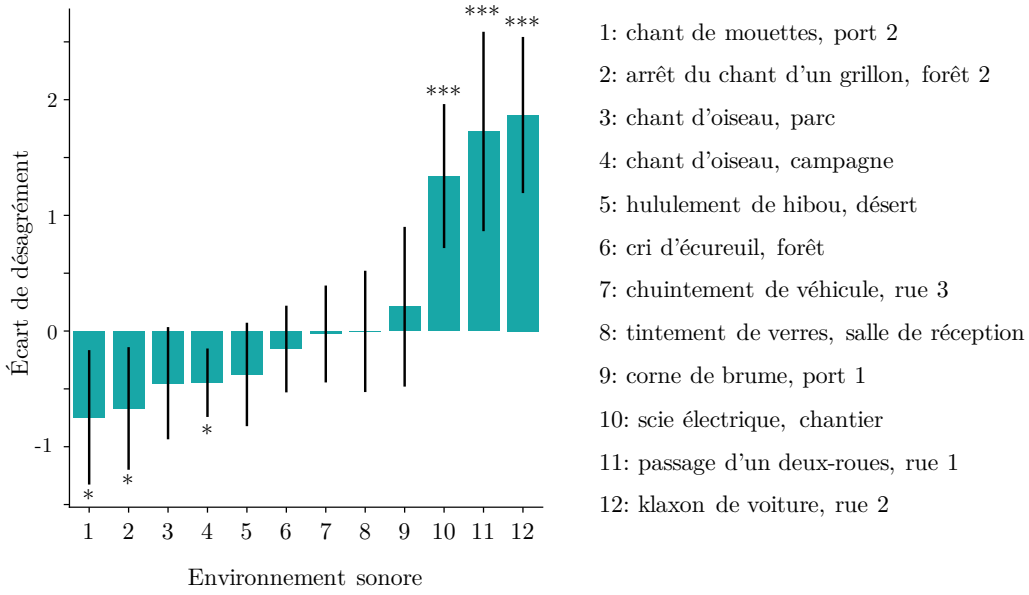

FIGURE 4.9 – Différence de désagrément moyennée sur deux secondes après l'évènement saillant dans chaque environnement, moyennée sur tous les participants. Les barres d'erreur représentent l'intervalle de confiance à 95%. La significativité des tests-t menés sur les distributions est indiquée par les étoiles (* : p<0,05, *** : p<0,01)

Dans les sources désagréables, on retient :

— la scie électrique sur le chantier (T(34) = 4,28, p < 0,001, p_{corr} < 0,001, cohen-d = 0,72) ;

— le passage d'un deux-roues dans la rue (T(32) = 3,9, p < 0,001, p_{corr} < 0,001, cohen-d = 0,68) ;

— le klaxon de voiture dans la rue (T(29) = 5,02, p < 0,001, p_{corr} < 0,001, cohen-d = 0,92) ;

On note que les effets des sources désagréables (effet statistique et effet sur l'évaluation de désagrément) sont plus forts que ceux des sources agréables.

On observe ainsi un lien entre la présence de sons saillants et l'évaluation du désagrément. Nous souhaiterions maintenant comprendre si cet effet est dû à une relation de causalité.

4.3.3 Analyse de causalité

4.3.3.1 Principe

Comme précisé en section 4.3.2, les évolutions temporelles de désagrément et de saillance sont largement corrélées, mais cette information ne suffit pas à comprendre les influences entre ces séries. Pour comprendre les influences mutuelles entre les séries temporelles, une analyse de causalité de Granger a donc été menée. Ce type d'analyse, initialement introduit en économétrie (Granger, 1969), suit le principe suivant.

Pour deux séries temporelles $A = (a_1, a_2, ..., a_n)$ et $B = (b_1, b_2, ..., b_n)$, on dit que A cause B au sens de Granger si on prédit mieux les valeurs présentes de B lorsqu'on prend en compte les valeurs passées de A en plus des valeurs passées de B. Autrement dit, A "Granger-cause" B si, en notant $\hat{b_t}$ la valeur prédite pour la série b à l'instant t :

$$E[(\hat{b_t} - b_t) \mid b_{t-1}, b_{t-2}, ...; a_{t-1}, a_{t-2}, ...]^2 < E[(\hat{b_t} - b_t) \mid b_{t-1}, b_{t-2}, ...]^2 \quad (4.1)$$

Plus précisément, une analyse de causalité de Granger prend en entrée un facteur de temps appelé "lag" et correspondant à la durée pendant laquelle on considère les valeurs des séries dans le passé pour prédire la valeur présente. D'autre part, nous ne connaissons pas le délai exact qui existe entre la perception d'un évènement sonore et la réaction physique répercutée sur le curseur. Nous devons donc inclure un paramètre, le "timeshift", correspondant au décalage temporel entre les deux séries temporelles à prendre en compte pour évaluer la causalité potentielle.

Si on note p le lag et ts le timeshift, alors A cause B avec un lag p et un timeshift ts, noté alors $A \xrightarrow{(p,ts)} B$, si :

$$E[(\hat{b_t} - b_t) \mid b_{t-1}, ..., b_{t-p}; a_{t-ts-1}, ..., a_{t-ts-p}]^2 < E[(\hat{b_t} - b_t) \mid b_{t-1}, ..., b_{t-p}]^2 \quad (4.2)$$

Un test de causalité de Granger teste l'hypothèse nulle selon laquelle il n'y a pas causalité. Si la probabilité renvoyée par le test, notée $p_{A \xrightarrow{(p,ts)} B}$, est inférieure au seuil de significativité α, alors $A \xrightarrow{(p,ts)} B$:

$$p_{A \xrightarrow{(p,ts)} B} < \alpha \qquad \Leftrightarrow \qquad A \xrightarrow{(p,ts)} B \tag{4.3}$$

On dit alors que A cause B, soit $A \to B$, si et seulement s'il existe une paire (p, ts) telle que $A \xrightarrow{(p,ts)} B$ mais pas $B \xrightarrow{(p,ts)} A$. Mathématiquement :

$$\exists \ (p, ts) \ / \begin{cases} A \xrightarrow{(p,ts)} B \\ \neg \ (B \xrightarrow{(p,ts)} A) \end{cases} \qquad \Leftrightarrow \qquad A \to B \tag{4.4}$$

Dans certains cas (si A et B dépendent d'une troisième grandeur par exemple), on pourrait avoir $A \to B$ mais aussi $B \to A$. Or, nous souhaitons voir si l'évolution d'une grandeur cause celle d'une autre, pas si les deux grandeurs sont entre-causées. Nous disons donc que A cause B de manière unidirectionnelle, noté $A \xrightarrow{UGC} B$ (UGC pour *Unidirectional Granger Causality*), si et seulement si $A \to B$ mais pas $B \to A$:

$$\begin{cases} A \to B \\ \neg \ (B \to A) \end{cases} \qquad \Leftrightarrow \qquad A \xrightarrow{UGC} B \tag{4.5}$$

Ce type d'analyse implique que les séries temporelles soient stationnaires (c'est-à-dire que leurs caractéristiques - moyenne, variance - ne varient pas dans le temps). Si ce n'est pas le cas, on réalise cette analyse sur leurs dérivées. Des tests de stationnarité (test augmenté de Dickey-Fuller) sont donc réalisés en amont sur toutes les séries temporelles considérées pour vérifier cette hypothèse.

4.3.3.2 Mise en oeuvre

Les scènes prises en compte dans cette analyse sont toutes les scènes contenant un évènement saillant dans le passage-clé pour chaque environnement. Les scènes d'une même paire étant égales en tout point sauf durant le passage-clé, nous retirons les scènes ne contenant pas d'élément saillant car elles ne contiennent pas d'information supplémentaire par rapport à leur scène appairée.

Notre analyse consiste à savoir si des variations de saillance causent de manière unidirectionnelle des variations dans l'évaluation de désagrément. Nous souhaitons également observer comment les variations d'autres caractéristiques (sonie, brillance, rugosité) causent également le désagrément. Nous ajoutons également le niveau pondéré A ($L_{A,eq}$) dans nos analyses, cette mesure étant réalisée comme dans Filipan et al. (2019) (niveau de pression acoustique avec pondération A et lissage sur une fenêtre glissante de 250 ms). Les tests de stationnarité montrent que les séries temporelles collectées ne sont majoritairement pas stationnaires (seules 33% le sont). En revanche, leurs dérivées sont toutes stationnaires. L'analyse de causalité porte donc sur les dérivées des séries temporelles.

Nous avons fixé la plage de variation des lags à [0 ; 500] ms en accord avec Filipan et al. (2019). Cette borne supérieure à 500 ms reflète les résultats de retards constatés dans les études d'imagerie cérébrale (Deshpande and Hu, 2012 ; Goebel et al., 2003). La plage de variation des timeshifts est établie à [0 ; 1500] ms, toujours en accord avec Filipan et al. (2019). Cette borne supérieure permet notamment de couvrir les temps de réaction à des stimuli sensoriels les plus élevés (Bigelow and Poremba, 2014). Du fait de l'échantillonnage des données, ces intervalles sont plus précisément [0 ; 512] ms et [0 ; 1540] ms.

Pour chaque scène, on teste la causalité de Granger entre la dérivée d'une caractéristique (saillance, brillance, rugosité, sonie) et celle du désagrément. Le niveau de significativité est corrigé (correction de Bonferroni)

Caractéristique	Causalités unidirectionnelles (%)
Saillance	83%
Sonie	58%
Brillance	58%
Rugosité	42%
L_{Aeq}	42%

TABLE 4.2 – Proportion des scènes pour lesquelles il y a causalité unidirectionnelle en fonction de la caractéristique issue du modèle de saillance

pour prendre en compte le grand nombre de tests réalisés dans l'espace des $(p, ts) \in ([0; 512], [0; 1540])$ (avec une fréquence d'échantillonnage de 15,6 Hz). On compte ensuite, pour chaque caractéristique, le nombre de scènes pour lesquelles il y a causalité unidirectionnelle sur le désagrément et on note la proportion correspondante. Les résultats sont présentés au tableau 4.2.

Dans 83% des scènes du corpus (10 scènes), une causalité unidirectionnelle est établie entre saillance prédite et désagrément. Cette proportion est de 58% (7 scènes) pour la sonie et la brillance, de 42% (5 scènes) pour la rugosité et le L_{Aeq}.

4.3.3.3 Modèles à Vecteur Autorégressif (VAR)

Lorsqu'il existe une relation de causalité entre deux séries, c'est-à-dire lorsque $A \rightarrow B$, nous cherchons tous les couples (p, ts) donnant lieu à $A \xrightarrow{(p,ts)} B$. Pour chacun de ces couples, un modèle à Vecteur Autorégressif (VAR) est implémenté, et le Critère d'Information Bayésien (BIC) est relevé. Nous cherchons le couple (p, ts) minimisant ce critère d'information. Autrement dit, nous cherchons le couple (p, ts) qui optimise le modèle VAR entre A et B dans l'espace des causalités unidirectionnelles.

Un modèle VAR à p lags entre deux grandeurs A et B avec un timeshift ts s'écrit comme suit :

$$\begin{cases} b_t = \beta_0 + \beta_{1,1} * b_{t-1} + ... + \beta_{1,p} * b_{t-p} + \beta_{2,1} * a_{t-ts-1} + ...+ \\ \qquad \beta_{2,p} * a_{t-ts-p} + \varepsilon_{1,t} \\ a_{t-ts} = \alpha_0 + \alpha_{1,1} * a_{t-ts-1} + ... + \alpha_{1,p} * a_{t-ts-p} + \alpha_{2,1} * b_{t-1} + ...+ \\ \qquad \alpha_{2,p} * b_{t-p} + \varepsilon_{2,t} \end{cases} \tag{4.6}$$

$\{\alpha_{i,j}, \beta_{i,j}\}_{i,j}$ sont les paramètres du modèle, $\varepsilon_{1,t}$ et $\varepsilon_{2,t}$ les erreurs. Le BIC permet de comparer entre eux les modèles optimisés avec les différentes valeurs de (p, ts) assurant une causalité unidirectionnelle : le modèle ayant le BIC le plus faible est considéré comme meilleur. Plus précisément, une différence de BIC inférieure à 2 constitue une preuve "faible" en faveur du modèle ayant le BIC le plus petit, une différence entre 2 et 6 une preuve "positive", une différence entre 6 et 10 une preuve "forte", une différence supérieure à 10 une preuve "très forte" (Raftery, 1995).

Nous pouvons observer les résultats des modèles VAR optimisés entre saillance et désagrément. Le couple (p, ts) qui optimise le modèle VAR entre ces deux grandeurs vaut, en moyenne sur les 10 scènes où il y a causalité unidirectionnelle :

— $p = 269 \pm 14$ ms
— $ts = 1,52 \pm 0,01$ s

Cette valeur de timeshift étant proche de la valeur maximale permise dans notre analyse (1,54 secondes), la même analyse a été réalisée en autorisant des timeshifts dans la plage de variation [0 ; 2] s pour éviter tout effet de plafonnement de cette valeur. On trouve alors un lag et un timeshift optimaux de :

— $p = 238 \pm 16$ ms
— $ts = 1,52 \pm 0,07$ s

Enfin, les valeurs de BIC obtenues pour les différents modèles entre les différentes caractéristiques (sonie, brillance, rugosité, $L_{A,eq}$) et le désagrément évalué confirment que la saillance permet de mieux prédire le désagrément que toutes les autres grandeurs. Le BIC du modèle optimisé entre saillance et désagrément est en effet toujours inférieur à celui de n'importe quel modèle optimisé entre une autre caractéristique et le désagrément, et la différence est toujours supérieure à 10. Il s'agit donc systématiquement d'une preuve "très forte" en faveur des modèles prenant en compte la saillance (Raftery, 1995). Ces résultats sont synthétisés en figure 4.10.

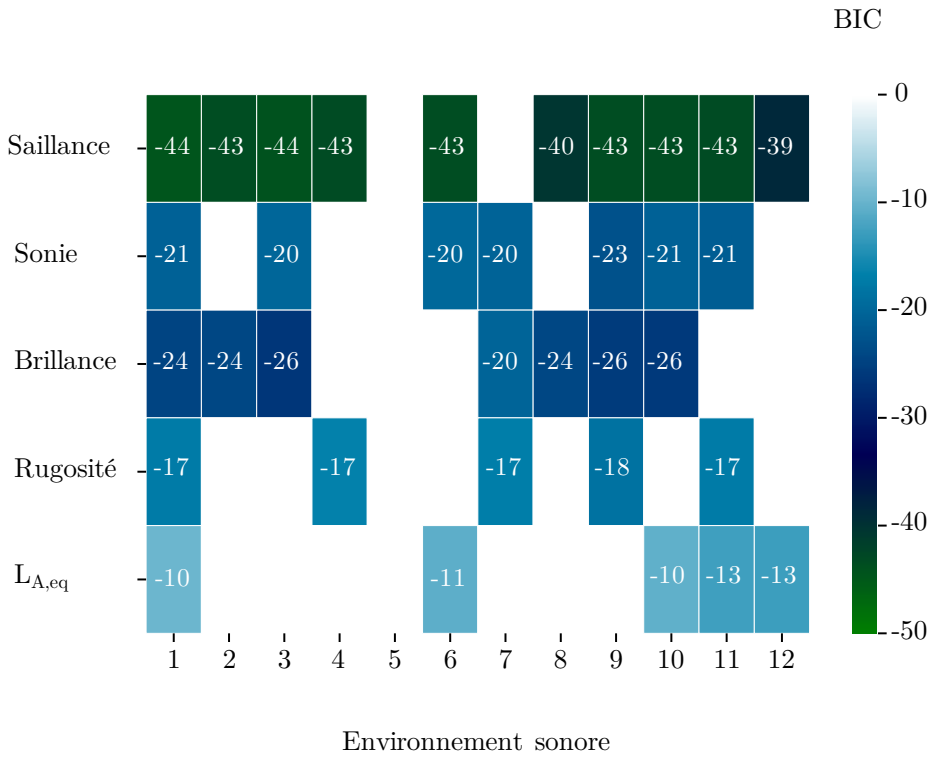

FIGURE 4.10 – Critère d'Information Bayésien (BIC) en fonction de la caractéristique entrée dans le modèle VAR avec le désagrément, par scène. Plus la valeur est faible, plus le modèle est bon. 1 : campagne, 2 : forêt 2, 3 : désert, 4 : réception, 5 : forêt, 6 : rue 2, 7 : rue 3, 8 : parc, 9 : rue 1, 10 : chantier, 11 : port 1, 12 : port 2

4.4 Discussion

Nous avons ainsi montré l'effet de la présence de sources saillantes sur le désagrément, avant d'établir formellement cette relation de causalité.

4.4.1 Effet de sources saillantes sur le désagrément

Les analyses réalisées sur les données collectées indiquent un effet de la saillance sur l'évaluation du désagrément sonore dans des scènes environnementales. De fait, la présence de certains éléments saillants dans les scènes sonores affecte significativement l'évaluation de désagrément par rapport à la même scène ne contenant pas cet élément. Nous avons notamment observé que les sources mécaniques impliquaient une hausse du désagrément perçu, alors que les sources naturelles étaient au contraire liées à une hausse de l'agrément. Ces résultats sont cohérents avec les résultats de la littérature. Ainsi, par exemple, Lavandier and Defréville (2006) ont observé que des sources sonores comme des voix et des oiseaux semblaient augmenter l'agrément des environnements sonores urbains, alors que des sources mécaniques (bus, cyclomoteurs, voitures) le détérioraient. Nilsson et al. (2007) ont également mis en évidence que la qualité de l'environnement sonore était fortement corrélée à l'identification des sons électro-mécaniques (corrélation négative) et à l'identification des sons naturels (corrélation positive).

L'effet des sources augmentant le désagrément semble par ailleurs être plus important que celui des sources augmentant l'agrément : les variations dans l'évaluation d'agrément ainsi que les tailles d'effet calculées sont plus faibles pour toutes les sources "agréables" que pour les sources "désagréables". Il est néanmoins difficile de conclure sur ce point car un effet de contraste avec le contexte sonore pourrait expliquer ces résultats autant qu'un effet plus marqué dans le sens désagréable que dans le sens agréable de manière générale. De fait, les sources relevées comme les plus désagréables (scie électrique, klaxon, passage du deux-roues) étaient présentes dans un

environnement par ailleurs assez calme : le chantier ne contenait que des impacts de marteau assez lointains dans un environnement urbain calme avant l'occurrence de la scie électrique, et les rues dans lesquelles le klaxon et le deux-roues émergeaient étaient assez calmes également, avec peu de bruit de circulation avant l'évènement saillant. Un évènement désagréable dans ces environnements relativement calmes pourrait donc causer plus de désagrément qu'un évènement agréable se produisant dans un environnement jugé comme déjà agréable avant l'évènement (un chant d'oiseau dans la campagne par exemple). D'autres études seraient nécessaires pour approfondir cette question du contraste et déterminer si les sources désagréables perturbent plus l'appréciation de l'environnement sonore que les sources agréables de manière générale.

4.4.2 Causes du désagrément

Il convient en premier lieu de noter que l'observation d'un lien entre deux grandeurs comme la saillance et le désagrément n'est pas immédiat. De fait, la saillance d'un évènement sonore est censée traduire la capacité de cet évènement à capturer l'attention d'un auditeur indépendamment de sa volonté. Le modèle de Huang et al. (2017) a de ce fait été entraîné sur des données issues d'un paradigme d'écoute dichotique, visant à assurer que les participants ne prêtaient pas spécifiquement attention à la scène évaluée.
Dans un paradigme de mesure du désagrément en continu, il est demandé aux participants d'évaluer à chaque instant à quel point l'environnement sonore leur paraît agréable/désagréable. Cela implique ainsi une écoute active de la scène, différente de celle mise en jeu dans une mesure de saillance. Les modes d'écoute différant pour la mesure des deux grandeurs considérées ici, l'observation du lien entre les deux pourrait être bruitée.

Malgré cela, les prédictions de saillance améliorent significativement la prédiction du désagrément dans la plupart des environnements sonores de notre corpus (10 sur 12), et ce malgré le bruit que pourrait engendrer les modes d'écoute différents impliqués dans les paradigmes de mesure d'agré-

ment et de mesure de saillance ayant permis de construire le modèle de prédiction (cf. section 4.4). La prédiction de saillance de Filipan et al. (2019) était moins systématiquement influente (49% des cas) dans leur étude (l'influence du niveau de pression acoustique est du même ordre dans les deux études : 38% dans leur cas, 42% dans le nôtre).

Cette différence pourrait être due à une meilleure évaluation de la saillance par le modèle de Huang et al. (2017) que par la modélisation réalisée dans l'étude de Filipan et al. (2019), le modèle de Huang et al. (2017) ayant notamment été entraîné sur des données perceptives. Par ailleurs, la nature des scènes pourrait également être à l'origine d'une différence. Comme précisé en section 4.1.3, les évaluations de désagrément et de saillance relevées par Filipan et al. (2019) étaient empreintes de l'effet des transitions entre différentes zones urbaines, et donc des variations importantes de ces grandeurs acoustiques à ces instants. En comparaison de ces variations, l'effet des sources sonores dans leur environnement pourrait donc avoir été amoindri. Dans notre corpus, chaque scène se déroulait intégralement dans un environnement sonore donné, et seule la présence de sources saillantes issues de cet environnement pouvait générer des variations de désagrément perçu. Enfin, les mesures de causalité utilisées dans les deux études ne sont pas exactement équivalentes : Filipan et al. (2019) ont en effet relevé la causalité entre prédictions de saillance et probabilité de changement dans les évaluations de désagrément (une mesure issue de la dérivée du désagrément) pour chaque scène et chaque participant. Le lissage issu du calcul de la moyenne du désagrément sur tous les participants dans notre étude pourrait minimiser l'impact d'évaluations irrégulières données par certains participants, et ainsi renforcer la mesure de causalité. La proportion de causalité obtenue dans nos travaux avec la mesure de niveau moyen d'énergie, similaire à la leur, semble cependant valider la comparaison de nos mesures respectives.

Nos résultats permettent également d'estimer le délai entre le moment où une source sonore émerge dans une scène et le moment où le désagrément perçu varie. Les résultats de nos analyses ont en effet montré que le délai optimisant les performances de nos modèles à vecteur auto-régressif entre

saillance et désagrément perçu était de $1,52 \pm 0,07$ s. Ces résultats sont cohérents avec les observations que l'on peut faire en figure 4.6, où l'on note un décalage temporel systématique entre les pics dans la courbe de saillance et ceux dans la courbe de désagrément. Ils sont également cohérents avec ceux de la littérature : Aumond et al. (2017b) ont relevé un délai d'environ 2 secondes entre le niveau sonore mesuré et l'évaluation de l'agrément, Kuwano and Namba (1985) un délai d'1 seconde entre le niveau sonore mesuré et une évaluation instantanée du niveau sonore perçu. Notre résultat, plus proche de celui de Aumond et al. (2017b) est cohérent avec le fait que l'évaluation du désagrément, plus complexe, implique des délais légèrement plus élevés que celle du niveau sonore.

On note que la saillance est un meilleur prédicteur de l'évaluation du désagrément que la sonie, elle-même étant meilleure que le $L_{A,eq}$. De même, les modèles auto-régressifs encodant la saillance sont toujours meilleurs que ceux encodant la sonie, eux même meilleurs que ceux encodant le $L_{A,eq}$. Ces résultats sont cohérents : la saillance est une mesure qui prend en compte les évolutions de la sonie, qui est elle-même obtenue à partir de mesures d'énergie par bande critique (Zwicker et al., 1991). Ces mesures semblent ainsi, depuis le $L_{A,eq}$ jusqu'à la saillance en passant par la sonie, pouvoir mieux caractériser les mécanismes cognitifs mis en jeu dans la perception des scènes sonores.

Les résultats d'analyse de causalité montrent également l'importance du timbre dans la perception de scènes sonores environnementales. Les variations de deux attributs du timbre (brillance et rugosité), par lesquels la capture attentionnelle se trouvait modulée au chapitre 2, causent en effet également des variations de désagrément, aussi souvent (dans autant de scènes) que la sonie pour la brillance, et toujours plus souvent (dans plus de scènes) et avec plus d'intensité que le $L_{A,eq}$. Or, Klein et al. (2015) ont montré que des indices caractérisant les modulations d'amplitude et le contenu dans les hautes fréquences améliorent, certes dans des proportions plus faibles que la sonie, les prédictions de jugements de gêne associée à des bruits de passages

de véhicules isolés. Par ailleurs, Hong and Jeon (2017) ont trouvé, dans des scènes sonores urbaines, que le niveau équivalent était la variable la plus influente pour expliquer des évaluations globales de désagrément. La comparaison de nos résultats avec ces résultats issus d'évaluations globales paraît néanmoins limitée : les indices utilisés pour juger un environnement dans sa globalité pourrait être différents de ceux influençant une évaluation continue guidée par la variabilité des événements sonores. Par ailleurs, l'importance particulière du timbre dans nos résultats peut s'expliquer par la nature de notre corpus qui ne se limite pas aux seuls environnements urbains. Il contient au contraire une quantité importante, pour la moitié des environnements, de sons issus de sources naturelles (chants d'oiseaux, insectes, écureuil, etc). Or, ces sources sont caractérisées par une sonie globalement plus faible et un contenu dans les hautes fréquences plus dense que les bruits de circulation notamment (Yang and Kang, 2013).

Nos résultats confirment ainsi que la saillance auditive est le meilleur indicateur pour aider à prédire l'agrément d'une scène sonore environnementale. En effet, ses variations causent (au sens de Granger) plus souvent les variations de désagrément perçu que tout autre descripteur étudié ici (sonie, brillance, rugosité et $L_{A,eq}$). De plus, lorsque les variations d'un autre descripteur causent également les variations de désagrément, la saillance permet une meilleure prédiction que ce descripteur (voir figure 4.10).
La saillance auditive semble ainsi être déterminante dans la perception de scènes sonores complexes, et semble plus pertinente que les autres indicateurs acoustiques ou psychoacoustiques, tels que la sonie ou le niveau $L_{A,eq}$. Ces résultats confirment ainsi l'importance d'étudier la saillance comme déterminant de la perception de scènes sonores, et ainsi de dépasser l'utilisation d'indicateurs s'appuyant uniquement sur des niveaux moyens d'énergie acoustique.

4.4.3 Implications pratiques et méthodologiques

L'évolution de la saillance telle que prédite par le modèle de Huang et al. (2017) cause les variations de désagrément perçu par les participants dans des scènes sonores environnementales variées. Par ailleurs, l'effet de sources sonores saillantes sur le désagrément dépend de la nature de ces sources, en accord avec les résultats de la littérature (Lavandier and Defréville, 2006 ; Nilsson et al., 2007). Les deux scènes pour lesquelles la causalité entre saillance et désagrément n'est pas significative sont d'ailleurs des scènes dans lesquelles l'évènement saillant n'a pas d'effet sur le désagrément (un cri d'écureuil dans la forêt et un chuintement de véhicule dans la rue). Ainsi, une prise en compte simultanée de la saillance et de la nature des sources présentes dans la scène (et donc de la direction de l'effet présumé sur le désagrément) pourrait constituer un bon prédicteur du désagrément sonore perçu dans une scène. Ce point pourrait déboucher sur des applications pratiques d'évaluation des environnements sonores, notamment d'environnements urbains afin d'en mesurer/prédire la qualité.

D'un point de vue méthodologique, nous avons proposé dans cette étude un protocole permettant d'établir l'influence d'une source sonore dans son environnement sur la perception de l'agrément sonore. En effet, la constitution de scènes sonores appairées, ne contenant qu'une différence d'un évènement sonore saillant, permet de comparer les évaluations en absence et en présence de la source saillante et d'en mesurer précisément l'effet, dans son contexte. Cette méthodologie n'a, à notre connaissance, pas encore été mise en oeuvre pour étudier l'impact de sources sonores sur la perception de l'environnement dans lequel elles se situent. Elle pourrait constituer un point de départ intéressant pour évaluer la hausse ou la baisse de désagrément engendrée par l'inclusion d'une source sonore dans un environnement donné. De fait, enrichir le paysage sonore avec des installations diffusant des sons visant à diminuer le désagrément est une approche de plus en plus considérée (Fraisse et al., 2021) ; cette méthodologie pourrait permettre d'en mesurer l'impact.

4.5 Conclusion

Nous avons ainsi mis en évidence le rôle de la saillance sur la perception du paysage sonore, notamment sur l'évaluation de l'un de ses attributs perceptifs principaux, le désagrément. Des variations de saillance prédite au cours du temps sont en effet responsables de variations de désagrément perçu, et ce de manière plus systématique qu'avec d'autres paramètres acoustiques ou psychoacoustiques couramment utilisés. En outre, la direction de l'effet des évènements saillants sur l'évaluation de désagrément dépend de la nature de la source, de manière cohérente avec les résultats connus de la littérature.

Nous avons ainsi confirmé que la présence d'évènements saillants module la perception d'une scène sonore (cf. chapitre 3), ici dans le cas de l'évaluation du désagrément dans des paysages sonores. Nous avons également confirmé le rôle du timbre sonore, notamment des deux attributs étudiés au chapitre 2, qui affecte la perception du désagrément autant que la sonie et plus qu'un niveau moyen d'énergie acoustique.

D'un point de vue méthodologique, le principe expérimental proposé dans cette étude permet de mettre en évidence l'effet d'une source sonore sur l'évaluation du désagrément de l'environnement au cours du temps. Cette méthodologie pourrait être reprise pour étudier l'impact de certaines sources sonores sur la perception de leur environnement, dans le cas d'études visant à améliorer l'expérience sonore urbaine ou à étudier l'impact de sources anthropiques dans des environnements naturels par exemple.

Nous apportons ainsi la confirmation théorique que la saillance est un indicateur pertinent pour évaluer la perception du paysage sonore. Nous recommandons donc l'utilisation de descripteurs basés sur une prédiction de la saillance dans le domaine de l'acoustique environnementale, en ce qu'ils sont de meilleurs prédicteurs du désagrément que les mesures de niveau sonore ou même de sonie.

Conclusion

Résultats principaux

Les travaux menés dans cette thèse ont permis d'apporter certaines réponses aux questionnements que nous relevions en introduction :
Comment certains sons parviennent-ils à s'imposer à nous ? Comment leur présence affecte notre perception et notre appréciation de l'environnement sonore ?

Nous avons ainsi mis en évidence au chapitre 2 que les propriétés de certains sons modulent directement notre attention de manière ascendante. La modulation d'un effet de capture attentionnelle par des propriétés sonores dans un paradigme psycchoacoustique a en effet permis de révéler la composante *stimulus-driven* de l'attention. Certaines propriétés de cette relation ont été révélées pour deux attributs du timbre en particulier, la brillance et la rugosité, ainsi confirmés comme déterminants de la saillance auditive.

De plus, l'étude menée au chapitre 3 a permis de montrer que la saillance modifie l'organisation de la perception de scènes sonores. Des sons rendus saillants localement dans des séquences mélodiques plus complexes qu'au chapitre précédent inversent en effet le phénomène de primauté du traitement holistique, et ce même pour des experts-musiciens déployant des mécanismes descendants visant à les ignorer.

Enfin, nous avons montré au chapitre 4 que les sources saillantes influencent l'appréciation de l'environnement sonore qui les englobe. En effet, les évènements saillants affectent significativement les évaluations de désagrément, et l'évolution de la saillance se trouve être le meilleur indicateur, parmi d'autres descripteurs acoustiques et psychoacoustiques, pour prédire le désagrément dans des scènes sonores de nos environnements.

Homme, Son, Environnement

Les travaux menés dans cette thèse permettent d'élargir, au moins partiellement, la compréhension que nous avons du rapport de l'homme à son environnement sonore.

Il semble ainsi que des éléments de cet environnement puissent s'imposer à nous indépendamment de notre volonté par le biais de certaines propriétés sonores et des mécanismes cognitifs associés. De plus, il apparaît que notre analyse d'une scène sonore, ainsi que notre perception et notre appréciation d'un environnement donné, sont affectées par la saillance de son contenu.

Notre appréciation de l'environnement résulte donc de l'effet des éléments saillants qui peuvent s'imposer à nous par le biais de leurs propriétés sonores. L'homme apparaît ainsi comme partiellement soumis au contenu de son environnement sonore, dont l'étude, la préservation, et l'amélioration devraient donc constituer des enjeux de développement majeur.

PERSPECTIVES

Au fil de cette thèse, un certain nombre de questions ont été soulevées et constituent des pistes de réflexion à mener dans la continuité directe de ces travaux. Nous en soulignons quelques unes ici.

L'étude menée au chapitre 2 valide une méthode expérimentale permettant de mettre en évidence l'effet de propriétés sonores sur l'effet de capture attentionnelle. Dans le paradigme proposé, nous pouvons observer la relation entre variations d'un paramètre acoustique et temps de capture attentionnelle. L'utilisation de cette méthode pour étudier tout un ensemble de caractéristiques, celles encodées par le modèle de Kothinti et al. (2021) par exemple, est donc envisageable. Cela pourrait permettre de comparer les influences respectives de chacune et ainsi enrichir les connaissances liées à l'influence du timbre sur la saillance. De telles études pourraient également permettre de répondre à la question de la similarité des lois liant la taille de l'effet de capture attentionnelle et les différentes caractéristiques étudiées.

Les résultats du chapitre 3 montrent que la présence de sons saillants affecte des principes propres à l'analyse de scènes auditives, dans ce cas la primauté du traitement holistique. Cet effet paraît persister malgré l'expertise de participants pourtant dotés de capacités pour adopter une écoute analytique plus performante. Une exploration plus détaillée des capacités des experts, à la manière de Susini et al. (2023), associée aux observations de leurs performances dans notre paradigme pourrait permettre de mieux comprendre par quel biais l'effet de la saillance pourrait être inhibé. Autrement dit, quelles composantes de l'expertise musicale pourraient renforcer l'inhibition de la capture attentionnelle ? De plus, les mécanismes étudiés pourraient varier, notamment en mettant en jeu une tâche impliquant la saillance dans des flux audio différents simultanés. Cela pourrait permettre de cible d'autres principes de l'analyse de scène auditive, la ségrégation de flux par exemple.

Enfin, les résultats du chapitre 4 établissent la saillance comme déterminant décisif de la perception d'environnements sonores, notamment de son appréciation. L'effet de sources saillantes sur des évaluations de désagrément est en effet significatif. L'influence de caractéristiques sonores en particulier pourrait être étudié, en mettant en oeuvre notre paradigme dans lequel les propriétés des sources seraient manipulées. Les caractéristiques à étudier pourraient être les mêmes que celles mentionnées ci-dessus, à explorer dans d'éventuels prolongements des travaux du chapitre 2. Ce genre d'étude pourrait mener à une cartographie dans l'espace (saillance, désagrément) de l'effet de différents attributs perceptifs. Par ailleurs, des incrustations de sources dont on souhaite étudier l'effet sur un paysage sonore en particulier (véhicules électriques sonifiés dans l'environnement urbain par exemple) permettraient, avec notre paradigme, d'évaluer, anticiper, et ainsi limiter d'éventuelles nuisances.

Annexes

Annexe 1 : Paradigme spatial du singleton additionnel

Les procédés nous permettant de localiser les sources sonores dans l'espace ont largement été étudiés (voir Moore (2012) pour une revue sur le sujet). L'"Interaural Time Difference" (ITD) et l'"Interaural Level Difference" (ILD) caractérisent respectivement la différence temporelle et la différence d'intensité avec lesquelles les deux oreilles reçoivent un même son. Ces deux indices sont utilisés pour localiser les sons en azimuth, l'un - l'ITD - plutôt à plus basse fréquence, l'autre - l'ILD - plutôt à plus haute fréquence (Rayleigh, 1907). La précision de la localisation d'une source dans l'espace peut être étudiée via la question de l'angle minimal audible, c'est-à-dire la plus petite séparation entre les angles des directions de deux sources à partir de laquelle une différence est souvent détectée (dans 75% des cas généralement - Savel (2009)).

Mills (1958) ont par exemple montré que la différence d'angle minimale perceptible pour la localisation d'une source sonore dans le plan azimutal variait en fonction de la fréquence de la source mais restait globalement inférieure à 10°. Klumpp and Eady (1956) ont mesuré des seuils de détection d'ITD de 9μs pour des bandes de bruit (contenant des fréquences entre 150 et 1700 Hz) et de 11 μs pour des tons purs sinusoïdaux à 1000 Hz. Cela correspond à des variations d'azimut d'environ 1°.

La précision angulaire de localisation semble ainsi suffisamment bonne pour faire la différence entre des sources diffusées depuis 5 haut-parleurs situés à 45° d'écart minimum les uns des autres dans le plan azimutal.

Notre première expérience pilote [5] visait ainsi à répondre à la question suivante : un participant peut-il percevoir 5 stimuli, répartis dans l'espace,

5. Note : à la suite d'un problème informatique, les données de cette expérience ont été perdues et nous sommes dans l'incapacité de présenter certaines informations avec précision (niveau de restitution, durée moyenne de l'expérience, détail des résultats notamment).

simultanément de manière distincte ?

Participants

6 personnes ont pris part à cette expérience (4 hommes, 2 femmes). Ils étaient tous consentants, âgés de 23 à 55 ans (moyenne : 32 ans) et ont tous fait part d'une audition normale. Ils n'ont pas reçu de rémunération.

Apparatus

L'expérience a été conçue et s'est déroulée sur un MacBook pro (2020), avec le logiciel Max (version 8) (https://cycling74.com/). Les stimuli ont été conçus avec le logiciel Max également. Ils sont présentés durant l'expérience grâce à des haut-parleurs Amadeus PMX 5A pilotés par une carte son Fireface 800 de la marque RME. Les haut-parleurs sont répartis dans le demi-plan azimutal avant, sur un demi-cercle de rayon 1 mètre et centré sur la tête du participant, aux angles : -90°, -45°, 0°, + 45°, + 90° (voir figure 11). L'expérience se déroule dans un studio du laboratoire STMS de l'Ircam.

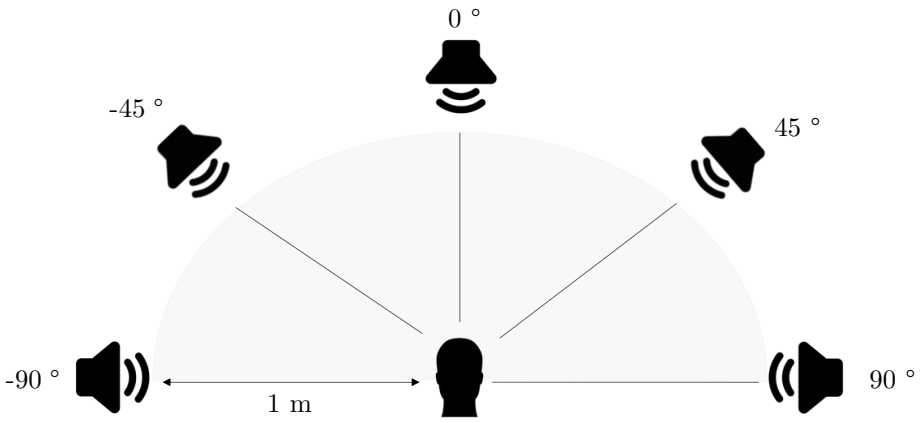

FIGURE 11 – Répartition des haut-parleurs dans le demi-plan azimutal avant pour le test pilote de perception de sources séparées dans l'espace. Les haut-parleurs sont placés à une distance de 1 m de la tête du participant, et tous les 45 degrés dans le demi-plan azimutal avant.

Stimuli

Différents types de stimuli sonores ont été testés, dans 4 sessions différentes pour cette expérience pilote :

1. Des tons purs sinusoïdaux : dans ce cas, chaque ton avait une fréquence différente. Plus précisément, 9 bandes critiques distinctes de l'oreille humaine (« Equivalent Rectangular Bandwidth ») entre 300 et 6000 Hz et leur fréquence centrale ont été retenues (Moore and Glasberg, 1983). La fréquence de chaque ton était tirée au hasard parmi une des 9 fréquences disponibles, et les différents sons diffusés simultanément à chaque essai avaient une fréquence différente. Ceci permettait de s'assurer que chaque son était situé dans une bande critique différente (Fletcher, 1940) ;
2. Des tons purs sinusoïdaux modulés en amplitude : dans ce cas, chaque ton avait une fréquence différente (parmi la même liste de fréquences possibles que pour les tons purs du cas 1) et une fréquence de modulation différente parmi 5 possibles. Plus précisément, les cinq fréquences de modulation d'amplitude possibles étaient inférieures au seuil à partir duquel la modulation d'amplitude est perçue comme un ajout de rugosité au son (modulations d'amplitudes entre $\sim$ 30 et 150 Hz - Arnal et al. (2015)). Le but était que les sons soient perçus comme des sons différents du cas 1, et non comme les mêmes sons rendus plus rugueux. Les fréquences de modulation d'amplitude étaient donc situées entre 3 et 20 Hz ;
3. Des bandes de bruit : dans ce cas, chaque bande de bruit blanc était filtrée autour d'une des fréquences (les mêmes que dans le cas des tons purs du cas 1), avec un facteur de qualité de 100 ;
4. Des échantillons d'instruments de musique : les instruments, issus de Wessel et al. (1987) étaient choisis dans l'espace des timbres de McAdams et al. (1995) pour avoir des timbres distincts. Les 5 instruments retenus étaient : hautbois/clavecin, piano, guitare pincé, violon frotté,

vibraphone (voir figure 12). Les cinq instruments jouaient la même note ;

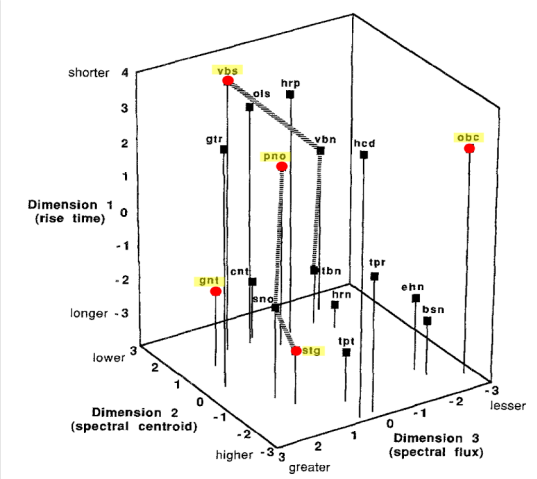

FIGURE 12 – Espace des timbres de McAdams et al. (1995). Les instruments surlignés en jaune et dont la position est pointée en rouge sont ceux utilisés dans le cas 4 de l'expérience pilote.

Les stimuli duraient tous 500 ms, avec une rampe en début et en fin de 20 ms.

Égalisation en sonie

Tous les stimuli étaient égalisés en sonie par le participant au début de l'expérience (par une méthode d'ajustement avec une cible de référence sur le haut-parleur central à une fréquence de 1000 Hz), et les différents haut-parleurs l'étaient également entre eux (par égalisation avec une cible à 1000 Hz sur le haut-parleur central).

Test de spatialisation préliminaire à l'expérience pilote

L'expérience pilote mettant en jeu la capacité des participants à percevoir des stimuli sonores de manière simultanée dans l'espace, un test de spatialisation a d'abord été mené. Le but était de voir si les participants étaient capables d'identifier, parmi les 5 haut-parleurs disposés comme dans l'expérience pilote,

lequel diffusait un son. De fait, pour que les participants entendent bien, dans l'expérience pilote, 5 sons issus de 5 directions différentes, il fallait que chaque stimulus puisse être entendu comme issu du bon endroit.
Chaque participant réalisait 60 essais. À chaque essai, un son était diffusé aléatoirement depuis l'un des cinq haut-parleurs. La procédure était une procédure de choix forcé à 5 alternatives : le participant devait dire depuis quel haut-parleur le son était diffusé.

Résultats du test préliminaire

Les participants ont identifié correctement le haut-parleur depuis lequel le son était émis dans 70 % des cas. L'erreur moyenne de localisation était de 23°, soit la moitié de la distance séparant deux haut-parleurs. La précision de localisation était donc suffisamment bonne pour mener l'expérience pilote.

Procédure de l'expérience pilote

Un essai se déroulait de la manière suivante : l'inscription « prêt ? » s'affichait à l'écran durant 500 millisecondes, puis l'inscription « Go ! » s'affichait et N sons étaient alors diffusés simultanément. N était égal à 3, 4 ou 5 de manière aléatoire d'un essai à l'autre. Les participants ne savaient pas jusqu'à combien de sons pouvaient être diffusées simultanément. La présentation des sons durait 500 ms. À la fin de la séquence, le participant devait dire combien de sons distincts il avait perçus (les réponses entre 1 et 5 étaient proposées). Lorsque le participant avait donné sa réponse, l'essai suivant était lancé. Une réponse était forcément requise pour passer à l'essai suivant.

Chaque participant passait 50 essais sur chaque type de stimuli (tons sinusoïdaux purs, tons sinusoïdaux purs modulés en amplitude, bandes de bruit, instruments).

Résultats

Quel que soit le type de stimuli utilisé, les participants ne sont jamais parvenus à percevoir plus de 3 sources simultanées. Aucune réponse au-delà de 3 n'a été donnée par les participants. Le taux de bonnes réponses (bon nombre de sources identifiées) oscillait ainsi entre 2% et 30 % selon les participants.

Discussion

La répartition des stimuli dans l'espace ne semble pas permettre une bonne séparation des items, car les participants ne distinguent jamais clairement au-delà de trois sons distincts. Ceci est problématique pour l'étude de la capture attentionnelle par un singleton additionnel car il faut être sûr que le participant perçoive bien 5 stimuli distinctement.

De plus, dans notre paradigme, tous les distracteurs devaient être identiques. Or, l'oreille humaine groupe naturellement les sons aux caractéristiques identiques pour former des flux audio (Bregman, 1994). Parmi les facteurs favorisant le groupement, on trouve entre autres la cohérence d'enveloppe, la synchronicité d'attaque, l'harmonicité, l'égalité des fréquences fondamentales. Le groupement des distracteurs, de même enveloppe, de même fréquence fondamentale, de même contenu spectral, en un seul flux, est donc un risque important. Ce groupement est incompatible avec la nécessaire capacité des participants à distinguer les items séparément les uns des autres (voir partie 2.1.1.2).

De plus, la précision de la localisation des sons dans l'espace peut dépendre des caractéristiques acoustiques de ces derniers, et notamment de leur fréquence (Klumpp and Eady, 1956 ; Mills, 1958 ; Stevens and Newman, 1936). Or, dans le paradigme du singleton additionnel que nous devions mettre en place, nous souhaitions pouvoir manipuler les caractéristiques acoustiques des items sonores pour tester leur influence sur l'effet de capture attentionnelle. Il était donc nécessaire de pouvoir manipuler ces caractéristiques librement sans que cela ne modifie la perception globale de la présentation des items.

Ce choix de ne pas retenir la présentation des items sonores dans l'espace s'est aussi imposé pour des questions pratiques liées aux conditions spécifiques

à la crise sanitaire que nous avons traversée en 2020-2021. La répartition spatiale des stimuli demande en effet un matériel et une installation plus imposants (utilisation de 5 haut-parleurs répartis dans l'espace). Une spatialisation des sons par utilisation d'HRTF ("Head-Related Transfer Function") aurait pu être envisagée pour tenter d'obtenir un rendu spatialisé sur un équipement mobile, mais la précision de la spatialisation grâce à cette méthode est sujette à de fortes variations individuelles (Blauert, 1997 ; Wenzel et al., 1993). Une répartition temporelle des stimuli permet de s'affranchir de ces problèmes, et de rendre l'expérience plus mobile : un ordinateur et un casque de laboratoire suffisent en effet pour une écoute dans une pièce calme. Des premiers tests pilotes ont ainsi pu être menés auprès de colocataires lors des périodes de confinement et de télétravail généralisé, dans des conditions proches de celles du laboratoire.

Annexe 2 : informations supplémentaires du chapitre 2

Brightness deviation (jnd)	1	1.9	4.6	8.3
p	0.14	0.04	<.001	<.001
$p_{adjusted}$	0.14	0.08	.001	<.001
cohen-d	0.24	0.41	0.93	0.97

TABLE 3 – Effect significance (p), adjusted significance ($p_{adjusted}$, with Holm corrections for repeating comparisons) and size (cohen-d) of t-tests on the response time increases depending on the different singleton brightness values (quantified in jnd) in experiment 2.

Roughness deviation (jnd)	1	2	5	10
p	0.07	0.001	0.003	<.001
$p_{adjusted}$	0.07	0.006	0.008	0.001
cohen-d	0.35	0.76	-	-

TABLE 4 – Effect significance, adjusted significance (Holm corrections for repeating comparisons) and size of t-tests on the response time increases depending on the presence of the different singleton roughness values (quantified in jnd) in experiment 3. The shaded columns correspond to conditions for which the distribution is not normal (revealed through a Shapiro test), and for which Wilcoxon tests were applied.

Brightness deviation (jnd)	-4	-2	2	4
p	<.001	0.02	0.03	<.001
$p_{adjusted}$	0.002	0.04	0.04	0.002
cohen-d	0.93	0.49	0.48	0.93

TABLE 5 – Effect significance, adjusted significance (Holm corrections for repeating comparisons) and size of t-tests on the response time increases depending on the different singleton brightness values (quantified in jnd) in experiment 4.

[Brightness, Roughness] deviation (jnd)	[2, 2]	[2, 5]	[5, 2]	[5, 5]
p	0.004	<.001	<.001	<.001
$p_{adjusted}$	0.01	0.004	0.002	<.001
cohen-d	0.67	-	0.95	-

TABLE 6 – Effect significance, adjusted significance (Holm corrections for repeating comparisons) and size of t-tests on the response time increases depending on the different singleton brightness values (quantified in jnd) in experiment 5. The shaded columns correspond to conditions for which the distribution is not normal (revealed through a Shapiro test), and for which Wilcoxon tests were applied.

Annexe 3 : questionnaire portant sur l'expertise musicale des participants experts-musiciens (chapitre 3)

Le questionnaire proposé était une version plus courte et adaptée du "Goldsmiths Musical Sophistication Index", ou "Gold-MSI", (Müllensiefen et al., 2014), excluant les deux parties du Gold-MS traitant des capacités de chant et des dimensions émotionnelles qui n'étaient pas pertinentes pour notre étude. Des questions générales ont été posées pour déterminer si les participants étaient droitiers ou gauchers, s'ils avaient l'oreille absolue, s'ils avaient un emploi lié à la musique ou au son, s'ils se considéraient comme des musiciens et s'ils étaient autodidactes ou s'ils avaient reçu une formation dans une institution musicale. Les trois parties principales du questionnaire étaient composées de plusieurs affirmations que les participants devaient évaluer sur une échelle de 0 à 6, de "Pas du tout d'accord" (0) à "Tout à fait d'accord" (6). La première partie évaluait les informations professionnelles des participants liées à la musique et au son (par exemple, "J'ai un travail lié au son") ; la deuxième partie évaluait les habitudes comportementales des participants en matière de musique (par exemple, "Je suis curieux des différents styles musicaux que je ne connais pas et je veux en savoir plus") ; la troisième partie était une auto-évaluation de la capacité d'écoute et d'audition de la musique (par exemple, "Je peux dire quand les gens chantent ou jouent en décalage par rapport au rythme"). Les résultats étaient ensuite normalisés et sommés pour obtenir une note moyenne par participant sur 10.

Les participants experts-musiciens présentaient un score moyen de 7,6 à ce questionnaire, avec un écart-type de 0,56. Le participant expert-musicien exclu des analyses car autodidacte et présentant des résultats particulièrement proches de ceux des non-musiciens a obtenu un score de 5,3. Son score dans la partie 3, portant sur les capacités d'écoute musicale, était de 2,67 (moyenne des experts-musiciens sur cette partie : 6,89).

Participant	EM1	EM2	EM3	EM4	EM5	EM6	EM7
Sexe	M	M	M	M	F	M	M
Age	26	55	51	65	23	38	32
Partie 0							
ID							
Je suis droitier · ère/ gaucher · ère	d	d	d	d	d	d	d
Je me considère comme un musicien	o	o	o	o	o	o	n
J'ai l'oreille absolue	n	n	n	n	n	n	n
J'ai un problème auditif	n	n	n	n	n	/	n
Q1	5	4	3	4	2	5	2
Q2	6	6	6	5	6	6	4
Q3	6	4	6	6	6	4	6
Q4	5	2	4	6	6	3	5
Moyenne Part 0 (sur 10)	9,17	6,67	7,92	8,75	8,33	7,50	7,08
Partie 1							
Q1	5	5	6	5	6	4	3
Q2	2	1	1	2	1	1	0
Q3	4	3	6	5	5	4	5
Q4	6	4	5	4	4	5	2
Q5	6	2	4	5	6	3	1
Q6	5	6	6	5	5	6	4
Q7	4	2	6	4	3	5	0
Moyenne Part 1 (sur 10)	7,62	5,48	8,10	7,14	7,14	6,67	3,57
Partie 2							
Q1	6	4	5	6	4	4	5
Q2	4	3	6	6	6	5	6
Q3	5	5	6	5	4	4	5
Q4	5	5	6	5	5	5	5
Q5	6	5	6	5	5	6	5
Q6	6	5	6	5	5	5	5
Q7	6	5	6	6	6	4	5
Q8	6	5	6	6	6	4	4
Q9	5	4	6	5	4	5	3
Moyenne Part 2 (sur 10)	9,07	7,59	9,81	9,07	8,33	7,78	7,96
Partie 3							
Je suis autodidacte		non	non	non	non	non	oui
Mon niveau de formation	cours particuliers + formation 1	DEM d'orgue et CEM érudition	Cours particuliers	Ecole de musique	école de musique fun 2ème cycle	CEM Jazz/ CNRR Marseille	
Je joue principale de cet instrument	chant	orgue	trompette	guitare	clarinette	trombone	basse
Je joue plutôt de la musique	pop-rock	classique	jazz	tout sauf techno	classique / pop rock	jazz	pop rock
question 1	3	3	3	2	3	2	2
question 2	3	6	6	6	6	6	0
question 3	2	3	5	6	6	6	0
question 4	6	6	6	6	6	6	6
question 5	2	3	1	2	1	2	0
Moyenne Part 3 (sur 10)	5,33	7,00	7,00	7,33	7,33	7,33	2,67
Moyenne (Part0)	9,167	6,667	7,917	8,750	8,333	7,500	7,083
Moyenne (Part1)	7,619	5,476	8,095	7,143	7,143	6,667	3,571
Moyenne (Part2)	9,074	7,593	9,815	9,074	8,333	7,778	7,963
Moyenne (Part3)	5,333	7,000	7,000	7,333	7,333	7,333	2,667
Moyenne	**7,80**	**6,68**	**8,21**	**8,08**	**7,79**	**7,32**	**5,32**

FIGURE 13 – Résultats des participants experts-musiciens au questionnaire.

Annexe 4 : boîtiers de réponse pour l'évaluation continue du désagrément (chapitre 4)

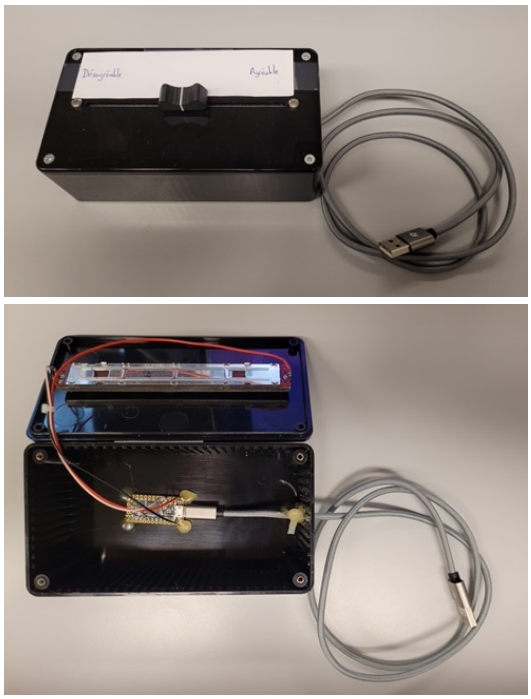

FIGURE 14 – Boîtiers utilisés dans l'expérience pour évaluer le désagrément en continu. Le signal est délivré au format midi (entiers de 0 à 127). La conception et la réalisation de ces contrôleurs a été menée en collaboration avec le Pôle Ingénierie et Prototype de l'IRCAM (E. Flety et A. Recher).

Annexe 5 : corrélations entre saillance et évaluations de désagrément (chapitre 4)

Scène	p-value
Campagne (a)	< 0,001
Campagne (s)	0,02
Forêt 1 (a)	< 0,001
Forêt 1 (s)	0,06
Forêt 2 (a)	< 0,001
Forêt 2 (s)	< 0,001
Désert (a)	< 0,001
Désert (s)	< 0,001
Parc (a)	< 0,001
Parc (s)	< 0,001
Salle de réception (a)	< 0,001
Salle de réception (s)	< 0,001
Rue 1 (a)	< 0,001
Rue 1 (s)	0,02
Rue 2 (a)	< 0,001
Rue 2 (s)	< 0,001
Rue 3 (a)	< 0,001
Rue 3 (s)	< 0,001
Chantier (a)	< 0,001
Chantier (s)	< 0,001
Port (a)	0,35
Port (s1)	< 0,001
Port (s2)	0,63

Table 7 – Résultat de significativité des tests de Pearson entre saillance et désagrément, pour chaque scène. Chaque environnement ayant donné lieu à deux scènes (l'une contenant un évènenement saillant dans le passage clé, l'autre non), la scène ne contenant pas d'évènement saillant dans le passage-clé est indiquée avec (a), celle en contenant un avec (s).

Annexe 6 : évaluations globales de désagrément (chapitre 4)

Dans une expérience supplémentaire, des participants ont évalué l'agrément global des scènes sonores diffusées dans l'expérience d'évaluation continue de désagrément. Dans le cas de cette expérience supplémentaire, les participants notaient chaque scène dans sa globalité sur la même échelle que celle utilisée dans l'expérience d'évaluation en continu.

Matériel et méthodes

Participants

38 participants âgés entre 18 et 35 ans (moyenne de 25 ans) ont volontairement pris part à cette expérience. Aucun n'a déclaré avoir des problèmes d'audition. Ils ont donné leur consentement par écrit avant l'expérience et ont été rémunérés pour leur participation.

Équipement et stimuli

Les stimuli étaient les mêmes en tout point que ceux diffusés dans l'expérience d'évaluation en continu (cf. 4.2.3). Ils étaient diffusés avec le même équipement (cf. 4.2.2). Les participants évaluaient le désagrément global sur la même interface, en manipulant cette fois le curseur sur l'écran avec la souris.

Procédure

La procédure était la même que pour l'expérience d'évaluation en continu, à un détail près : les participants ne devaient pas évaluer l'agrément sonore en continu, mais simplement attendre la fin de chaque scène sonore pour donner une note de désagrément relative à la scène sur une échelle continue allant de 0 à 10, de la même manière que réalisé par (Aumond et al., 2017b).

Résultats et discussion

Les évaluations de désagrément global pour chaque scène et chaque participant ont été relevées. On peut ainsi observer l'évaluation moyenne de désagrément global de chaque scène en figure 15. On note bien d'une part les environnements sonores naturels jugés comme les plus agréables, et d'autre part les environnements sonores urbains jugés plus désagréables. La moyenne de ces désagrément moyens est de 4,5, ce qui confirme le bon équilibre du corpus de scènes.

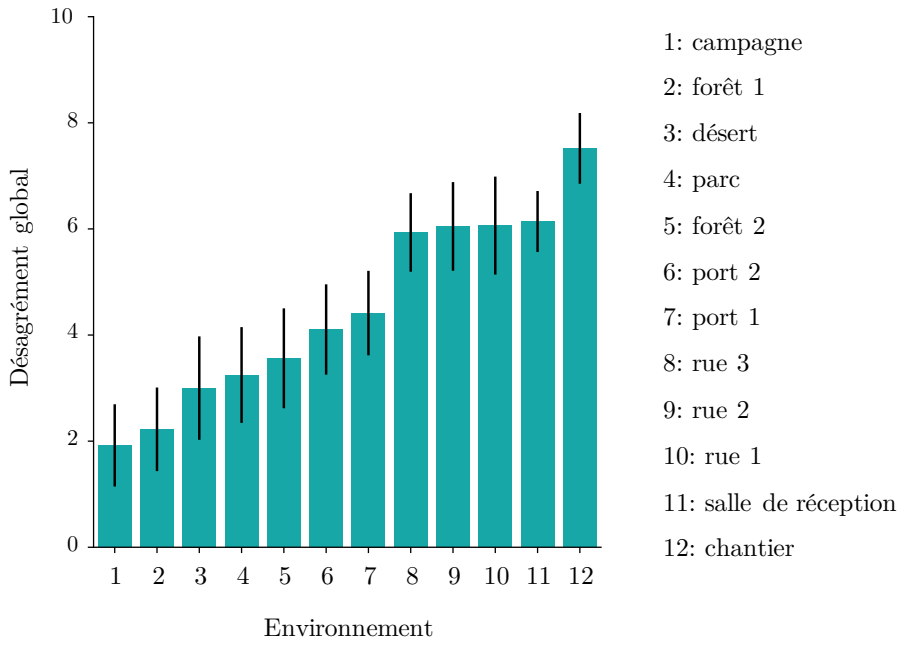

FIGURE 15 – Évaluation globale de désagrément sonore des différents environnements sonores du corpus (évaluation sur les scènes comportant l'évènement saillant).

On peut également observer en figure 15 l'écart de désagrément, pour chaque environnement sonore, entre la scène contenant un évènement saillant et la scène n'en contenant pas. On note alors que l'effet de la présence d'un évènement saillant dans le passage-clé des scènes concernées n'est pas significatif sur les évaluations globales de désagrément. Ce résultat pourrait être du à l'effet des évènements saillants situés plus tardivement dans la scène qui

pourrait influencer le jugement global de manière plus marquée. De fait, on sait que le jugement global d'une scène peut être sujet à un effet de récence : la fin de scène influence davantage le jugement global que le reste (Västfjäll, 2004). De plus, les observations faites au chapitre 4, notamment l'évolution comparée des désagréments et des saillances de chaque paire de scènes (voir figure 4.7 pour un exemple), semblent montrer que l'écart d'évaluation du désagrément dû à un premier évènement saillant se réduit significativement au moment où un deuxième évènement saillant survient.

Ces résultats invitent à mener des études complémentaires : on pourrait notamment comparer des évaluations de désagrément dans des scènes sonores contenant un évènement saillant dans le passage-clé mais pas d'autre évènement saillant ensuite. On observerait alors l'effet de chaque source isolée sur la scène dans sa globalité.

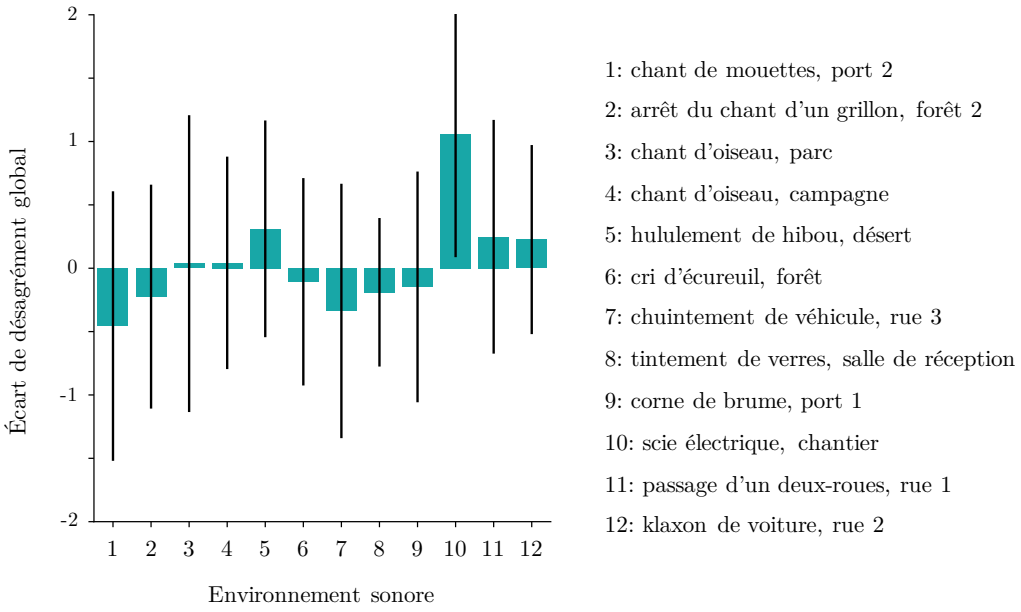

FIGURE 16 – Différences d'évaluation globale de désagrément sonore entre la scène contenant l'évènement saillant et la scène n'en contenant pas pour les différents environnements sonores du corpus.

Les barres d'erreur représentent les intervalles de confiance à 95

Collaborations et productions scientifiques

Cette thèse a été l'occasion de collaborations et de travaux variés, parfois au-delà du sujet de la saillance auditive qui est au coeur de cette étude. Certains de ces travaux ont mené à des publications que je mentionne ici. Je tiens à remercier ici les personnes avec lesquelles j'ai pris plaisir à collaborer et à explorer de nouveaux concepts, parfois par simple curiosité. Grâce à elles, j'ai eu la chance de pouvoir travailler et apprendre dans des champs disciplinaires plus variés que je ne l'aurais imaginé au début de cette thèse.

Bouvier, B., Susini, P., Marquis-Favre, C., & Misdariis, N. (2023). Revealing the stimulus-driven component of attention through modulations of auditory salience by timbre attributes. *Scientific Reports*, 13(1), 6842. (Bouvier et al., 2023b)

Bouvier, B., Susini, P., Marquis-Favre, C., & Misdariis, N. (2023). Auditory salience : A study of the influence of timbre attributes using the additional singleton paradigm. *Acta Amazonica*. (Bouvier et al., 2023a)

Bouvier, B., Ponsot, E., & Susini, P. (2023). Reorganization of temporal local/global processing with auditory salience, *10th Convention of the European Acoustics Association (Forum Acusticum 2023)* (Bouvier et al., 2023c).

Frid, E., Pauletto, S., Bouvier, B., & Fraticelli, M. (2023). A Dual-Task Experimental Methodology for Exploration of Saliency of Auditory Notifications in a Retail Soundscape. In *International Conference on Auditory Displays-Sonification for the Masses (ICAD 2023), Norrköping, Sweden, 26-30 June 2023*. (Frid et al., 2023).

Loiseau, R., Bouvier, B., Teytaut, Y., Vincent, E., Aubry, M., & Landrieu, L. (2022). A model you can hear : Audio identification with playable prototypes. *arXiv preprint arXiv :2208.03311*. (Loiseau et al., 2022)

Teytaut, Y., Bouvier, B., & Roebel, A. (2022, September). A study on constraining Connectionist Temporal Classification for temporal audio alignment. In *Interspeech 2022* (pp. 5015-5019). ISCA. (Teytaut et al., 2022)

Références

Abel, S. M. (2015). Duration discrimination of noise and tone bursts. (September 1971) :1219–1223.

ADEME (2021). Le coût social du bruit en france.

Alayrac, M., Marquis-Favre, C., and Viollon, S. (2011). Total annoyance from an industrial noise source with a main spectral component combined with a background noise. *The Journal of the Acoustical Society of America*, 130(1) :189–199.

Aletta, F., Kang, J., and Axelsson, Ö. (2016). Soundscape descriptors and a conceptual framework for developing predictive soundscape models. *Landscape and Urban Planning*, 149 :65–74.

Aletta, F., Oberman, T., Mitchell, A., Kang, J., and Consortium, S. (2023). Preliminary results of the soundscape attributes translation project (satp) : lessons learned and next steps. In *Proc. Forum Acusticum*.

Allen, E. J. and Oxenham, A. J. (2014). Symmetric interactions and interference between pitch and timbre. *The Journal of the Acoustical Society of America*, 135(3) :1371–1379.

Allport, D. A., Antonis, B., and Reynolds, P. (1972). On the division of attention : A disproof of the single channel hypothesis. *Quarterly journal of experimental psychology*, 24(2) :225–235.

ANSES (2013). Évaluation des impacts sanitaires extra-auditifs du bruit environnemental : avis de l'anses : Rapport d'expertise collective.

Arnal, L. H., Flinker, A., Kleinschmidt, A., Giraud, A. L., and Poeppel, D. (2015). Human Screams Occupy a Privileged Niche in the Communication Soundscape. *Current Biology*, 25(15) :2051–2056.

Arnal, L. H., Kleinschmidt, A., Spinelli, L., Giraud, A. L., and Mégevand, P. (2019). The rough sound of salience enhances aversion through neural synchronisation. *Nature Communications*, 10(1) :1–12.

Aumond, P., Can, A., De Coensel, B., Botteldooren, D., Ribeiro, C., and Lavandier, C. (2017a). Modeling soundscape pleasantness using perceptual assessments and acoustic measurements along paths in urban context. *Acta Acustica united with Acustica*, 103(3) :430–443.

Aumond, P., Can, A., De Coensel, B., Ribeiro, C., Botteldooren, D., and Lavandier, C. (2017b). Global and continuous pleasantness estimation of the soundscape perceived during walking trips through urban environments. *Applied Sciences*, 7(2) :144.

Axelsson, Ã., Nilsson, M. E., and Berglund, B. (2012). The swedish soundscape-quality protocol. *The Journal of the Acoustical Society of America*, 131(4) :3476–3476.

Axelsson, Ö., Nilsson, M. E., and Berglund, B. (2010). A principal components model of soundscape perception. *The Journal of the Acoustical Society of America*, 128(5) :2836–2846.

Belopolsky, A. V. and Theeuwes, J. (2010). No capture outside the attentional window. *Vision research*, 50(23) :2543–2550.

Belopolsky, A. V., Zwaan, L., Theeuwes, J., and Kramer, A. F. (2007). The size of an attentional window modulates attentional capture by color singletons. *Psychonomic bulletin & review*, 14(5) :934–938.

Berglund, B. (1998). Community noise in a public health perspective. In *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, volume 1998, pages 1683–1688. Institute of Noise Control Engineering.

Berglund, B., Hassmén, P., and Preis, A. (2002). Annoyance and spectral contrast are cues for similarity and preference of sounds. *Journal of Sound and vibration*, 250(1) :53–64.

Bever, T. G. and Chiarello, R. J. (1974). Cerebral dominance in musicians and nonmusicians. *Science*, 185(4150) :537–539.

Bey, C. and McAdams, S. (2002). Schema-based processing in auditory scene analysis. *Perception & psychophysics*, 64 :844–854.

Bey, C. and McAdams, S. (2003). Postrecognition of interleaved melodies as an indirect measure of auditory stream formation. *Journal of experimental psychology : human perception and performance*, 29(2) :267.

Bidet-Caulet, A., Bottemanne, L., Fonteneau, C., Giard, M.-H., and Bertrand, O. (2015). Brain dynamics of distractibility : interaction between top-down and bottom-up mechanisms of auditory attention. *Brain topography*, 28 :423–436.

Bigelow, J. and Poremba, A. (2014). Achilles' ear ? inferior human short-term and recognition memory in the auditory modality. *PloS one*, 9(2) :e89914.

Bizley, J. K. and Cohen, Y. E. (2013). The what, where and how of auditory-object perception. *Nature Reviews Neuroscience*, 14(10) :693–707.

Black, E., Stevenson, J. L., and Bish, J. P. (2017). The role of musical experience in hemispheric lateralization of global and local auditory processing. *Perception*, 46(8) :956–975.

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

Bockstael, A., De Coensel, B., Lercher, P., and Botteldooren, D. (2011). Influence of temporal structure of the sonic environment on annoyance. In *10th International Congress on Noise as a Public Health Problem (ICBEN-2011)*, volume 33, pages 945–952. Institute of Acoustics.

Bogdanov, V., Marquis-Favre, C., Cottet, M., Beffara, B., Perrin, F., Dumortier, D., and Ellermeier, W. (2022). Nature and the city : Audiovisual interactions in pleasantness and psychophysiological reactions. *Applied Acoustics*, 193 :108762.

Boswijk, V., Loerts, H., and Hilton, N. H. (2020). Salience is in the eye of the beholder : Increased pupil size reflects acoustically salient variables. *Ampersand*, 7 :100061.

Botteldooren, D., Andringa, T., Aspuru, I., Brown, A. L., Dubois, D., Guastavino, C., Kang, J., Lavandier, C., Nilsson, M., Preis, A., et al. (2015). From sonic environment to soundscape. *Soundscape and the built environment*, 36 :17–42.

Bouvet, L., Rousset, S., Valdois, S., and Donnadieu, S. (2011). Global precedence effect in audition and vision : Evidence for similar cognitive styles across modalities. *Acta psychologica*, 138(2) :329–335.

Bouvier, B., Susini, P., Marquis-Favre, C., and Misdariis, N. (2023a). Auditory salience : A study of the influence of timbre attributes using the additional singleton paradigm. *Acta Amazonica*.

Bouvier, B., Susini, P., Marquis-Favre, C., and Misdariis, N. (2023b). Revealing the stimulus-driven component of attention through modulations of auditory salience by timbre attributes. *Scientific Reports*, 13(1) :6842.

Bouvier, B., Susini, P., and Ponsot, E. (2023c). Reorganization of temporal local/global processing with auditory salience. *10th Convention of the European Acoustics Association (Forum Acusticum 2023)*.

Bregman, A. S. (1994). *Auditory scene analysis : The perceptual organization of sound*. MIT press.

Broadbent, D. E. (1958). *Perception and communication*. Elsevier.

Cain, R., Jennings, P., and Poxon, J. (2013). The development and application of the emotional dimensions of a soundscape. *Applied acoustics*, 74(2) :232–239.

CGEDD (2017). Réflexion prospective sur une politique de réduction des nuisances sonores.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America*, 25(5) :975–979.

Cichy, R. M. and Teng, S. (2017). Resolving the neural dynamics of visual and auditory scene processing in the human brain : a methodological approach. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 372(1714) :20160108.

Colle, H. A. and Welsh, A. (1976). Acoustic masking in primary memory. *Journal of verbal learning and verbal behavior*, 15(1) :17–31.

Dalton, P. and Fraenkel, N. (2012). Gorillas we have missed : Sustained inattentional deafness for dynamic events. *Cognition*, 124(3) :367–372.

Dalton, P. and Hughes, R. W. (2014). Auditory attentional capture : Implicit and explicit approaches. *Psychological Research*, 78(3) :313–320.

Dalton, P. and Lavie, N. (2004). Auditory Attentional Capture : Effects of Singleton Distractor Sounds. *Journal of Experimental Psychology : Human Perception and Performance*, 30(1) :180–193.

Dalton, P. and Lavie, N. (2007). Overriding auditory attentional capture. *Perception and Psychophysics*, 69(2) :162–171.

de Cheveigné, A. (1999). Waveform interactions and the segregation of concurrent vowels. *The Journal of the Acoustical Society of America*, 106(5) :2959–2972.

De Coensel, B. and Botteldooren, D. (2010). A model of saliency-based auditory attention to environmental sound. *20th International Congress on Acoustics 2010, ICA 2010 - Incorporating Proceedings of the 2010 Annual Conference of the Australian Acoustical Society*, 5(August) :3480–3487.

Dehais, F., Causse, M., Vachon, F., Régis, N., Menant, E., and Tremblay, S. (2014). Failure to detect critical auditory alerts in the cockpit : evidence for inattentional deafness. *Human factors*, 56(4) :631–644.

Deshpande, G. and Hu, X. (2012). Investigating effective brain connectivity from fmri data : past findings and current issues with reference to granger causality analysis. *Brain connectivity*, 2(5) :235–245.

Desimone, R. and Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1) :193–222.

Deutsch, J. A. and Deutsch, D. (1963). Attention : Some theoretical considerations. *Psychological review*, 70(1) :80.

Deutsch, J. A., Deutsch, D., Lindsay, P., and Treisman, A. M. (1967). Comments on "selective attention : Perception or response ?" reply. *Quarterly Journal of Experimental Psychology*, 19(4) :362–367.

Devergie, A., Grimault, N., Tillmann, B., and Berthommier, F. (2010). Effect of rhythmic attention on the segregation of interleaved melodies. *The Journal of the Acoustical Society of America*, 128(1) :EL1–EL7.

Dowling, W. J., Lung, K. M.-T., and Herrbold, S. (1987). Aiming attention in pitch and time in the perception of interleaved melodies. *Perception & Psychophysics*, 41 :642–656.

Driver, J. (2001). A selective review of selective attention research from the past century. *British Journal of Psychology*, 92(1) :53–78.

Duangudom, V. and Anderson, D. V. (2007). Using auditory saliency to understand complex auditory scenes. *European Signal Processing Conference*, (Eusipco) :1206–1210.

Ellermeier, W. and Hellbrück, J. (1998). Is Level Irrelevant in "Irrelevant Speech" ? Effects of Loudness, Signal-to-Noise Ratio, and Binaural Unmasking. *Journal of Experimental Psychology : Human Perception and Performance*, 24(5) :1406–1414.

Ellermeier, W. and Zimmer, K. (1997). Individual differences in susceptibility to the “irrelevant speech effect”. *The Journal of the Acoustical Society of America*, 102(4) :2191–2199.

Engle, R. W. (2002). Working memory capacity as executive attention. *Current directions in psychological science*, 11(1) :19–23.

Fastl, H. (1991). Evaluation and measurement of perceived average loudness. In *Contributions to psychological acoustics*. Bibliotheks-und Informationssystem der Univ. Oldenburg.

Fiebig, A. and Sottek, R. (2015). Contribution of peak events to overall loudness. *Acta Acustica united with Acustica*, 101(6) :1116–1129.

Filipan, K., De Coensel, B., Aumond, P., Can, A., Lavandier, C., and Botteldooren, D. (2019). Auditory sensory saliency as a better predictor of change than sound amplitude in pleasantness assessment of reproduced urban soundscapes. *Building and Environment*, 148 :730–741.

Fletcher, H. (1940). Auditory patterns. *Reviews of modern physics*, 12(1) :47.

Fletcher, H. and Munson, W. A. (1933). Loudness, its definition, measurement and calculation. *Bell System Technical Journal*, 12(4) :377–430.

Folk, C. L. and Remington, R. (1998). Selectivity in distraction by irrelevant featural singletons : evidence for two forms of attentional capture. *Journal of Experimental Psychology : Human perception and performance*, 24(3) :847.

Folk, C. L., Remington, R. W., and Johnston, J. C. (1992). Involuntary covert orienting is contingent on attentional control settings. *Journal of Experimental Psychology : Human perception and performance*, 18(4) :1030.

Folk, C. L., Remington, R. W., and Wright, J. H. (1994). The structure of attentional control : contingent attentional capture by apparent motion, abrupt onset, and color. *Journal of Experimental Psychology : Human perception and performance*, 20(2) :317.

Fraisse, V., Wanderley, M. M., and Guastavino, C. (2021). Comprehensive framework for describing interactive sound installations : Highlighting trends through a systematic review. *Multimodal Technologies and Interaction*, 5(4) :19.

Frid, E., Pauletto, S., Bouvier, B., and Fraticelli, M. (2023). A dual-task experimental methodology for exploration of saliency of auditory notifications in a retail soundscape. In *International Conference on Auditory Displays-Sonification for the Masses (ICAD 2023), Norrköping, Sweden, 26-30 June 2023*.

Garrido, M. I., Kilner, J. M., Stephan, K. E., and Friston, K. J. (2009). The mismatch negativity : a review of underlying mechanisms. *Clinical neurophysiology*, 120(3) :453–463.

Gaspelin, N. and Luck, S. J. (2018). The role of inhibition in avoiding distraction by salient stimuli. *Trends in cognitive sciences*, 22(1) :79–92.

Gatersleben, B. and Griffin, I. (2017). Environmental stress. *Handbook of environmental psychology and quality of life research*, pages 469–485.

Gibson, B. S. and Kelsey, E. M. (1998). Stimulus-driven attentional capture is contingent on attentional set for displaywide visual features. *Journal of Experimental Psychology : Human perception and performance*, 24(3) :699.

Goebel, R., Roebroeck, A., Kim, D.-S., and Formisano, E. (2003). Investigating directed cortical interactions in time-resolved fmri data using vector autoregressive modeling and granger causality mapping. *Magnetic resonance imaging*, 21(10) :1251–1261.

Gozalo, G. R., Carmona, J. T., Morillas, J. B., Vílchez-Gómez, R., and Escobar, V. G. (2015). Relationship between objective acoustic indices and subjective assessments for the quality of soundscapes. *Applied Acoustics*, 97 :1–10.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica : journal of the Econometric Society*, pages 424–438.

Grimault, N. (2013). Rapport d'habilitation à diriger des recherches. Centre de Recherche en Neurosciences de Lyon, Université Lyon 1.

Guastavino, C. (2006). The ideal urban soundscape : Investigating the sound quality of french cities. *Acta Acustica united with Acustica*, 92(6) :945–951.

Gygi, B. and Shafiro, V. (2011). The incongruency advantage for environmental sounds presented in natural auditory scenes. *Journal of Experimental Psychology : Human Perception and Performance*, 37(2) :551.

Herholz, S. C. and Zatorre, R. J. (2012). Musical training as a framework for brain plasticity : behavior, function, and structure. *Neuron*, 76(3) :486–502.

Herranz-Pascual, K., Aspuru, I., and García, I. (2010). Proposed conceptual model of environmental experience as framework to study the soundscape. In *Inter Noise*, pages 2904–2912.

Hillstrom, A. P. and Yantis, S. (1994). Visual motion and attentional capture. *Perception & psychophysics*, 55(4) :399–411.

Hoffman, J. E. and Subramaniam, B. (1995). The role of visual attention in saccadic eye movements. *Perception & psychophysics*, 57(6) :787–795.

Hong, J. Y. and Jeon, J. Y. (2017). Relationship between spatiotemporal variability of soundscape and urban morphology in a multifunctional urban area : A case study in seoul, korea. *Building and Environment*, 126 :382–395.

Huang, N., Elhilali, M., Kaya, E. M., Elhilali, M., Xiang, J., Shamma, S. A., Simon, J. Z., Kaya, E. M., and Elhilali, M. (2017). Auditory salience using natural soundscapes. *Frontiers in Human Neuroscience*, 7(MAY) :1–15.

IFOP (2014). Sondage ifop de septembre 2014 sur un échantillon de 1001 personnes, représentatif de la population française âgée de 18 ans et plus.

ISO (2014). Acoustics–soundscape–part 1 : definition and conceptual framework (iso 12913-1 : 2014).

ISO (2019). Acoustics–soundscape–part 3 : Data analysis (iso 12913–2 :2018).

Itatani, N. and Klump, G. M. (2017). Animal models for auditory streaming. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 372(1714) :20160112.

Itti, L. and Koch, C. (2001). Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3) :194–203.

James, W. (1890). *The principles of psychology*, volume 1. Cosimo, Inc.

JNA (2016). Enquête ifop auprès d'un échantillon de 1003 personnes, représentatif de la population française âgée de 15 ans et plus.

Jones, D. M. and Macken, W. J. (1993). Irrelevant Tones Produce an Irrelevant Speech Effect : Implications for Phonological Coding in Working Memory. *Journal of Experimental Psychology : Learning, Memory, and Cognition*, 19(2) :369–381.

Jonides, J. and Yantis, S. (1988). Uniqueness of abrupt visual onset in capturing attention. *Perception & psychophysics*, 43(4) :346–354.

Justus, T. and List, A. (2005). Auditory attention to frequency and time : an analogy to visual local–global stimuli. *Cognition*, 98(1) :31–51.

Kalinli, O. and Narayanan, S. (2007). A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech. *International Speech Communication Association - 8th Annual Conference of the International Speech Communication Association, Interspeech 2007*, 4 :2452–2455.

Kane, M. J. and Engle, R. W. (2003). Working-memory capacity and the control of attention : the contributions of goal neglect, response competition, and task set to stroop interference. *Journal of experimental psychology : General*, 132(1) :47.

Kang, J. (2006). *Urban sound environment*. CRC Press.

Kang, J. and Schulte-Fortkamp, B. (2018). *Soundscape and the built environment*. CRC press.

Kaya, E. M. and Elhilali, M. (2012). A temporal saliency map for modeling auditory attention. *2012 46th Annual Conference on Information Sciences and Systems, CISS 2012*.

Kaya, E. M. and Elhilali, M. (2014). Investigating bottom-up auditory attention. *Frontiers in Human Neuroscience*, 8(MAY) :1–12.

Kaya, E. M. and Elhilali, M. (2017). Modelling auditory attention. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 372(1714).

Kaya, E. M., Huang, N., and Elhilali, M. (2020). Pitch, Timbre and Intensity Interdependently Modulate Neural Responses to Salient Sounds. *Neuroscience*, 440 :1–14.

Kayser, C., Petkov, C. I., Lippert, M., and Logothetis, N. K. (2005). Mechanisms for allocating auditory attention : An auditory saliency map. *Current Biology*, 15(21) :1943–1947.

Kim, K., Lin, K. H., Walther, D. B., Hasegawa-Johnson, M. A., and Huang, T. S. (2014). Automatic detection of auditory salience with optimized linear filters derived from human annotation. *Pattern Recognition Letters*, 38(1) :78–85.

Kimchi, R. (1992). Primacy of wholistic processing and global/local paradigm : a critical review. *Psychological bulletin*, 112(1) :24.

Kirmse, U., Jacobsen, T., and Schröger, E. (2009). Familiarity affects environmental sound processing outside the focus of attention : An event-related potential study. *Clinical neurophysiology*, 120(5) :887–896.

Klein, A., Marquis-Favre, C., Weber, R., and Trollé, A. (2015). Spectral and modulation indices for annoyance-relevant features of urban road single-vehicle pass-by noises. *The Journal of the Acoustical Society of America*, 137(3) :1238–1250.

Klumpp, R. and Eady, H. (1956). Some measurements of interaural time difference thresholds. *The Journal of the Acoustical Society of America*, 28(5) :859–860.

Koffka, K. (1935). *Principles of Gestalt psychology*, volume 44. Routledge.

Köhler, W. (1967). Gestalt psychology. *Psychologische Forschung*, 31(1) :XVIII–XXX.

Kollmeier, B., Brand, T., and Meyer, B. (2008). Perception of speech and sound. In *Springer handbook of speech processing*, pages 61–82. Springer.

Kondo, H. M. and Kashino, M. (2009). Involvement of the thalamocortical loop in the spontaneous switching of percepts in auditory streaming. *Journal of Neuroscience*, 29(40) :12695–12701.

Kondo, H. M., van Loon, A. M., Kawahara, J.-I., and Moore, B. C. (2017). Auditory and visual scene analysis : an overview. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 372(1714) :20160099.

Koreimann, S., Gula, B., and Vitouch, O. (2014). Inattentional deafness in music. *Psychological research*, 78(3) :304–312.

Kothinti, S. R., Huang, N., and Elhilali, M. (2021). Auditory salience using natural scenes : An online study. *The Journal of the Acoustical Society of America*, 150(4) :2952–2966.

Kubovy, M. (2017). Concurrent-pitch segregation and the theory of indispensable attributes. In *Perceptual organization*, pages 55–98. Routledge.

Kubovy, M. and Van Valkenburg, D. (2001). Auditory and visual objects. *Cognition*, 80(1-2) :97–126.

Kuwano, S. and Namba, S. (1985). Continuous judgment of level-fluctuating sounds and the relationship between overall loudness and instantaneous loudness. *Psychological research*, 47(1) :27–37.

LaBerge, D. (1983). Spatial extent of attention to letters and words. *Journal of Experimental Psychology : Human Perception and Performance*, 9(3) :371.

Lachter, J., Forster, K. I., and Ruthruff, E. (2004). Forty-five years after broadbent (1958) : still no identification without attention. *Psychological review*, 111(4) :880.

Lauter, J. L., Herscovitch, P., Formby, C., and Raichle, M. E. (1985). Tonotopic organization in human auditory cortex revealed by positron emission tomography. *Hearing research*, 20(3) :199–205.

Lavandier, C., Aumond, P., Can, A., Gontier, F., Lagrange, M., and Petit, G. (2021). Urban sensor network for characterizing the sound environment in lorient (france) through an automatic assessment of traffic, voice and bird presence ratios. In *European Congress on Noise Control Engineering (EuroNoise)*.

Lavandier, C. and Defréville, B. (2006). The contribution of sound source characteristics in the assessment of urban soundscapes. *Acta acustica united with Acustica*, 92(6) :912–921.

Lemaitre, G., Susini, P., Winsberg, S., McAdams, S., and Letinturier, B. (2007). The sound quality of car horns : a psychoacoustical study of timbre. *Acta acustica united with Acustica*, 93(3) :457–468.

Lercher, P. (1998). Deviant dose-response curves for traffic noise in'sensitive areas'? In *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, volume 1998, pages 710–713. Institute of Noise Control Engineering.

Liao, H.-I., Kidani, S., Yoneya, M., Kashino, M., and Furukawa, S. (2016). Correspondences among pupillary dilation response, subjective salience of sounds, and loudness. *Psychonomic bulletin & review*, 23 :412–425.

Loiseau, R., Bouvier, B., Teytaut, Y., Vincent, E., Aubry, M., and Landrieu, L. (2022). A model you can hear : Audio identification with playable prototypes. *arXiv preprint arXiv :2208.03311*.

Luck, S. J., Gaspelin, N., Folk, C. L., Remington, R. W., and Theeuwes, J. (2021). Progress toward resolving the attentional capture debate. *Visual Cognition*, 29(1) :1–21.

Mack, A. and Rock, I. (1998a). *Inattentional Blindness*. MIT Press.

Mack, A. and Rock, I. (1998b). Inattentional blindness : Perception without attention. *Visual attention*, 8 :55–76.

Maquestiaux, F. (2017). *Psychologie de l'attention*, volume 1. De Boeck Supérieur.

Marinato, G. and Baldauf, D. (2019). Object-based attention in complex, naturalistic auditory streams. *Scientific reports*, 9(1) :2854.

Marozeau, J., de Cheveigné, A., McAdams, S., and Winsberg, S. (2003). The dependency of timbre on fundamental frequency. *The Journal of the Acoustical Society of America*, 114(5) :2946–2957.

Marr, D. (1982). *Vision : A computational investigation into the human representation and processing of visual information*. MIT press.

McAdams, S. (2019). The perceptual representation of timbre. *Timbre : Acoustics, perception, and cognition*, pages 23–57.

McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., and Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres : Common dimensions, specificities, and latent subject classes. *Psychological Research*, 58(3) :177–192.

McDermott, J. H. (2009). The cocktail party problem. *Current Biology*, 19(22) :R1024–R1027.

Melara, R. D. and Marks, L. E. (1990). Interaction among auditory dimensions : Timbre, pitch, and loudness. *Perception & psychophysics*, 48(2) :169–178.

Merzenich, M. M., Colwell, S. A., and Andersen, R. A. (1982). Auditory forebrain organization. In *Cortical sensory organization*, pages 43–57. Springer.

Mevorach, C., Humphreys, G. W., and Shalev, L. (2006). Opposite biases in salience-based selection for the left and right posterior parietal cortex. *Nature neuroscience*, 9(6) :740–742.

Mietlicki, C., Mietlicki, F., Ribeiro, C., Gaudibert, P., and Vincent, B. (2014). The harmonica project, new tools to assess environmental noise and better inform the public. In *Proceedings of the forum acusticum conference, Krakow, Poland*, pages 7–12.

Miles, C., Jones, D. M., and Madden, C. A. (1991). Locus of the irrelevant speech effect in short-term memory. *Journal of Experimental Psychology : Learning, Memory, and Cognition*, 17(3) :578.

Miller, J. (1991). Reaction time analysis with outlier exclusion : Bias varies with sample size. *The quarterly journal of experimental psychology*, 43(4) :907–912.

Mills, A. W. (1958). On the minimum audible angle. 237(1958).

Moore, B. C. (2012). *An introduction to the psychology of hearing*. Brill.

Moore, B. C. and Glasberg, B. R. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *The journal of the acoustical society of America*, 74(3) :750–753.

Moore, B. C. and Gockel, H. (2002). Factors influencing sequential stream segregation. *Acta Acustica United with Acustica*, 88(3) :320–333.

Moray, N. (1959). Attention in dichotic listening : Affective cues and the influence of instructions. *Quarterly journal of experimental psychology*, 11(1) :56–60.

Morel, J., Marquis-Favre, C., Viollon, S., and Alayrac, M. (2012). A laboratory study on total noise annoyance due to combined industrial noises. *Acta Acustica united with Acustica*, 98(2) :286–300.

Most, S. B., Simons, D. J., Scholl, B. J., and Chabris, C. F. (2000). Sustained inattentional blindness. *Psyche*, 6(14).

Müllensiefen, D., Gingras, B., Musil, J., and Stewart, L. (2014). The musicality of non-musicians : An index for assessing musical sophistication in the general population. *PloS one*, 9(2) :e89642.

Murphy, S., Spence, C., and Dalton, P. (2017). Auditory perceptual load : A review. *Hearing Research*, 352 :40–48.

Navon, D. (1977). Forest before trees : The precedence of global features in visual perception. *Cognitive psychology*, 9(3) :353–383.

Neisser, U. and Becklen, R. (1975). Selective looking : Attending to visually specified events. *Cognitive psychology*, 7(4) :480–494.

Newby, E. A. and Rock, I. (1998). Inattentional blindness as a function of proximity to the focus of attention. *Perception*, 27(9) :1025–1040.

Nilsson, M., Botteldooren, D., and De Coensel, B. (2007). Acoustic indicators of soundscape quality and noise annoyance in outdoor urban areas. In *Proceedings of the 19th International Congress on Acoustics*.

Nilsson, M. E. and Berglund, B. (2006). Soundscape quality in suburban green areas and city parks. *Acta Acustica united with Acustica*, 92(6) :903–911.

Noorden, L. P. A. S. (1975). *Temporal coherence in the perception of tone sequences*. PhD thesis, VAM.

Oldoni, D., De Coensel, B., Bockstael, A., Boes, M., De Baets, B., and Botteldooren, D. (2015). The acoustic summary as a tool for representing urban sound environments. *Landscape and Urban Planning*, 144 :34–48.

OMS (2011). *Burden of disease from environmental noise : Quantification of healthy life years lost in Europe*. World Health Organization. Regional Office for Europe.

Ouimet, T., Foster, N. E., and Hyde, K. L. (2012). Auditory global-local processing : Effects of attention and musical experience. *The Journal of the Acoustical Society of America*, 132(4) :2536–2544.

Pantev, C., Hoke, M., Lehnertz, K., Lütkenhöner, B., Anogianakis, G., and Wittkowski, W. (1988). Tonotopic organization of the human auditory cortex revealed by transient auditory evoked magnetic fields. *Electroencephalography and clinical neurophysiology*, 69(2) :160–170.

Pashler, H. (1988). Cross-dimensional interaction and texture segregation. *Perception & psychophysics*, 43(4) :307–318.

Pitt, M. A. (1994). Perception of pitch and timbre by musically trained and untrained listeners. *Journal of experimental psychology : human perception and performance*, 20(5) :976.

Pomerantz, J. R. (2017). Perceptual organization in information processing. In *Perceptual organization*, pages 141–180. Routledge.

Ponsot, E., Susini, P., Saint Pierre, G., and Meunier, S. (2013). Temporal loudness weights for sounds with increasing and decreasing intensity profiles. *The Journal of the Acoustical Society of America*, 134(4) :EL321–EL326. Publisher : Acoustical Society of America.

Powell, C. A. (1979). *A summation and inhibition model of annoyance response to multiple community noise sources*. National Aeronautics and Space Administration.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological methodology*, pages 111–163.

Rayleigh, L. (1907). Xii. on our perception of sound direction. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 13(74) :214–232.

Remington, R. W., Johnston, J. C., and Yantis, S. (1992). Involuntary attentional capture by abrupt onsets. *Perception & Psychophysics*, 51(3) :279–290.

Ricciardi, P., Delaitre, P., Lavandier, C., Torchia, F., and Aumond, P. (2015). Sound quality indicators for urban places in paris cross-validated by milan data. *The Journal of the Acoustical Society of America*, 138(4) :2337–2348.

Robertson, L. C. (1986). From gestalt to neo-gestalt. *Approaches to cognition : Contrasts and controversies*, pages 159–188.

Rodriguez-Hidalgo, A., Pelaez-Moreno, C., and Gallardo-Antolin, A. (2018). The Robustness of Echoic Log-Surprise Auditory Saliency Detection. *IEEE Access*, 6 :72083–72093.

Romani, G. L., Williamson, S. J., and Kaufman, L. (1982). Tonotopic organization of the human auditory cortex. *Science*, 216(4552) :1339–1340.

Salame, P. and Baddeley, A. (1982). Disruption of short-term memory by unattended speech : Implications for the structure of working memory. *Journal of verbal learning and verbal behavior*, 21(2) :150–164.

Salamé, P. and Baddeley, A. (1989). Effects of Background Music on Phonological Short-term Memory. *The Quarterly Journal of Experimental Psychology Section A*, 41(1) :107–122.

Sanders, L. D. and Poeppel, D. (2007). Local and global auditory processing : Behavioral and erp evidence. *Neuropsychologia*, 45(6) :1172–1186.

Savel, S. (2009). Individual differences and left/right asymmetries in auditory space perception. i. localization of low-frequency sounds in free field. *Hearing research*, 255(1-2) :142–154.

Schafer, R. (1977). *The Tuning of the World*. Knopf.

Schauerte, B. and Stiefelhagen, R. (2013). Wow !' Bayesian surprise for salient acoustic event detection. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 6402–6406.

Shepherd, M., Findlay, J. M., and Hockey, R. J. (1986). The relationship between eye movements and spatial attention. *The Quarterly Journal of Experimental Psychology*, 38(3) :475–491.

Simons, D. J. (2000). Attentional capture and inattentional blindness. *Trends in cognitive sciences*, 4(4) :147–155.

Simons, D. J. and Chabris, C. F. (1999). Gorillas in our midst : Sustained inattentional blindness for dynamic events. *perception*, 28(9) :1059–1074.

Snyder, J. S., Gregg, M. K., Weintraub, D. M., and Alain, C. (2012). Attention, awareness, and the perception of auditory scenes. *Frontiers in psychology*, 3 :15.

Sohlberg, M. M. and Mateer, C. A. (1987). Effectiveness of an attention-training program. *Journal of clinical and experimental neuropsychology*, 9(2) :117–130.

Southwell, R., Baumann, A., Gal, C., Barascud, N., Friston, K., and Chait, M. (2017). Is predictability salient ? A study of attentional capture by auditory patterns. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 372(1714).

Southworth, M. F. (1967). *The sonic environment of cities.* PhD thesis, Massachusetts Institute of Technology.

Stevens, S. S. and Newman, E. B. (1936). The localization of actual sources of sound. *The American journal of psychology*, 48(2) :297–306.

Stilwell, B. T. and Gaspelin, N. (2021). Attentional suppression of highly salient color singletons. *Journal of Experimental Psychology : Human Perception and Performance*, 47(10) :1313.

Susini, P., Jiaouan, S. J., Brunet, E., Houix, O., and Ponsot, E. (2020). Auditory local–global temporal processing : evidence for perceptual reorganization with musical expertise. *Scientific Reports*, 10(1) :16390.

Susini, P., McAdams, S., and Smith, B. K. (2002). Global and continuous loudness estimation of time-varying levels. *Acta Acustica united with Acustica*, 88(4) :536–548.

Susini, P., Wenzel, N., Houix, O., and Ponsot, E. (2023). Psychophysical characterization of auditory temporal and frequency streaming capacities for listeners with different levels of musical expertise. *JASA Express Letters*, 3(8).

Talamini, F., Altoè, G., Carretti, B., and Grassi, M. (2017). Musicians have better memory than nonmusicians : A meta-analysis. *PloS one*, 12(10) :e0186773.

Talavage, T. M., Sereno, M. I., Melcher, J. R., Ledden, P. J., Rosen, B. R., and Dale, A. M. (2004). Tonotopic organization in human auditory

cortex revealed by progressions of frequency sensitivity. *Journal of neurophysiology*, 91(3) :1282–1296.

Tardieu, J., Misdariis, N., Langlois, S., Gaillard, P., and Lemercier, C. (2015). Sonification of in-vehicle interface reduces gaze movements under dual-task condition. *Applied ergonomics*, 50 :41–49.

Tarlao, C., Aumond, P., Lavandier, C., and Guastavino, C. (2023). Converging towards a french translation of soundscape attributes : Insights from quebec and france. *Applied Acoustics*, 211 :109572.

Teytaut, Y., Bouvier, B., and Roebel, A. (2022). A study on constraining connectionist temporal classification for temporal audio alignment. In *Interspeech 2022*, pages 5015–5019. ISCA.

Theeuwes, J. (1992). Perceptual selectivity for color and form. *Perception Psychophysics*, 51(6) :599–606.

Theeuwes, J. (1993). Visual selective attention : A theoretical analysis. *Acta psychologica*, 83(2) :93–154.

Theeuwes, J. (1994). Stimulus-driven capture and attentional set : selective search for color and visual abrupt onsets. *Journal of Experimental Psychology : Human perception and performance*, 20(4) :799.

Theeuwes, J. (2010). Top–down and bottom–up control of visual selection. *Acta psychologica*, 135(2) :77–99.

Theeuwes, J., Kramer, A. F., Hahn, S., and Irwin, D. E. (1998). Our eyes do not always go where we want them to go : Capture of the eyes by new objects. *Psychological Science*, 9(5) :379–385.

Titchener, E. B. (1909). The experimental psychology of thought.

Tkacz-Domb, S. and Yeshurun, Y. (2018). The size of the attentional window when measured by the pupillary response to light. *Scientific reports*, 8(1) :1–7.

Tordini, F., Bregman, A. S., and Cooperstock, J. R. (2013). Toward an improved model of auditory saliency. *Icad*, pages 189–196.

Tordini, F., Bregman, A. S., and Cooperstock, J. R. (2016). Prioritizing foreground selection of natural chirp sounds by tempo and spectral centroid. *Journal on Multimodal User Interfaces*, 10 :221–234.

Treisman, A. (1986). Properties, parts, and objects.

Treisman, A. and Geffen, G. (1967). Selective attention : Perception or response ? *Quarterly Journal of Experimental Psychology*, 19(1) :1–17.

Treisman, A. M. (1960). Contextual cues in selective listening. *Quarterly Journal of Experimental Psychology*, 12(4) :242–248.

Treisman, A. M. (1964). Verbal cues, language, and meaning in selective attention. *The American journal of psychology*, 77(2) :206–219.

Treisman, A. M. and Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12(1) :97–136.

Treisman, A. M. and Riley, J. G. (1969). Is selective attention selective perception or selective response ? a further test. *Journal of Experimental Psychology*, 79(1p1) :27.

Tremblay, S. and Jones, D. M. (1999). Change of intensity fails to produce an irrelevant sound effect : Implications for the representation of unattended sound. *Journal of Experimental Psychology : Human Perception and Performance*, 25(4) :1005–1015.

Treue, S. (2003). Visual attention : the where, what, how and why of saliency. *Current opinion in neurobiology*, 13(4) :428–432.

Truax, B. (1984). *Acoustic communication*. Ablex Publishing Corporation.

Trudeau, C., Steele, D., Dumoulin, R., and Guastavino, C. (2018). Sounds in the city : differences in urban noise management strategies across cities. *Proceedings of InterNOISE 2018*.

Tsiami, A., Katsamanis, A., Maragos, P., and Vatakis, A. (2016). Towards a behaviorally-validated computational audiovisual saliency model. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2847–2851. IEEE.

Tsuchida, T. and Cottrell, G. (2012). Auditory saliency using natural statistics. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 34.

Underwood, G. (1974). Moray vs. the rest : The effects of extended shadowing practice. *The Quarterly journal of experimental psychology*, 26(3) :368–372.

Uttal, W. R. (1988). *On seeing forms*. Psychology Press.

van Noorden, L. P. A. S. (1975). Temporal Coherence in the Perception of Tone Sequences, Institute for Perception. *Institute for Perception Research*, Ph. D.(1975).

Van Renterghem, T. and Botteldooren, D. (2016). View on outdoor vegetation reduces noise annoyance for dwellers near busy roads. *Landscape and urban planning*, 148 :203–215.

Vandierendonck, A. (2017). A comparison of methods to combine speed and accuracy measures of performance : A rejoinder on the binning procedure. *Behavior research methods*, 49(2) :653–673.

Vannier, M., Misdariis, N., Susini, P., and Grimault, N. (2018). How does the perceptual organization of a multi-tone mixture interact with partial and global loudness judgments ? *The Journal of the Acoustical Society of America*, 143(1) :575–593.

Västfjäll, D. (2004). The "end effect" in retrospective sound quality evaluation. *Acoustical science and technology*, 25(2) :170–172.

Veale, R., Hafed, Z. M., and Yoshida, M. (2017). How is visual salience computed in the brain ? insights from behaviour, neurobiology and modelling. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 372(1714) :20160113.

Vuilleumier, P. (2005). How brains beware : neural mechanisms of emotional attention. *Trends in cognitive sciences*, 9(12) :585–594.

Warner, C. B., Juola, J. F., and Koshino, H. (1990). Voluntary allocation versus automatic capture of visual attention. *Perception & psychophysics*, 48(3) :243–251.

Waye, K. P. and Öhrström, E. (2002). Psycho-acoustic characters of relevance for annoyance of wind turbine noise. *Journal of sound and vibration*, 250(1) :65–73.

Wenhart, T. and Altenmüller, E. (2019). A tendency towards details ? inconsistent results on auditory and visual local-to-global processing in absolute pitch musicians. *Frontiers in psychology*, 10 :31.

Wenzel, E. M., Arruda, M., Kistler, D. J., and Wightman, F. L. (1993). Localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America*, 94(1) :111–123.

Wertheimer, M. (1938). Gestalt theory.

Wessel, D., Bristow, D., and Settel, Z. (1987). Control of phrasing and articulation in synthesis. In *ICMC*.

Winkler, I. (2007). Interpreting the mismatch negativity. *Journal of psychophysiology*, 21(3-4) :147–163.

Woldorff, M. G., Gallen, C. C., Hampson, S. A., Hillyard, S. A., Pantev, C., Sobel, D., and Bloom, F. E. (1993). Modulation of early sensory processing in human auditory cortex during auditory selective attention. *Proceedings of the National Academy of Sciences*, 90(18) :8722–8726.

Wood, N. and Cowan, N. (1995). The cocktail party phenomenon revisited : how frequent are attention shifts to one's name in an irrelevant auditory channel ? *Journal of Experimental Psychology : Learning, Memory, and Cognition*, 21(1) :255.

Wundt, W. M. (1893). *Grundzüge der physiologischen Psychologie*, volume 2. W. Engelmann.

Yang, M. and Kang, J. (2013). Psychoacoustical evaluation of natural and urban sounds in soundscapes. *The Journal of the Acoustical Society of America*, 134(1) :840–851.

Yang, W. and Kang, J. (2005). Soundscape and sound preferences in urban squares : a case study in sheffield. *Journal of urban design*, 10(1) :61–80.

Yantis, S. (2000). Goal-directed and stimulus-driven determinants of attentional control. *Attention and Performance*, 18 :71–103.

Yantis, S. and Hillstrom, A. P. (1994). Stimulus-driven attentional capture : evidence from equiluminant visual objects. *Journal of experimental psychology : Human perception and performance*, 20(1) :95.

Yantis, S. and Jonides, J. (1984). Abrupt visual onsets and selective attention : evidence from visual search. *Journal of Experimental Psychology : Human perception and performance*, 10(5) :601.

Zhao, S., Yum, N. W., Benjamin, L., Benhamou, E., Yoneya, M., Furukawa, S., Dick, F., Slaney, M., and Chait, M. (2019). Rapid ocular responses are modulated by bottom-up-driven auditory salience. *Journal of Neuroscience*, 39(39) :7703–7714.

Zwicker, E. and Fastl, H. (2013). *Psychoacoustics : Facts and models*, volume 22. Springer Science & Business Media.

Zwicker, E., Fastl, H., Widmann, U., Kurakata, K., Kuwano, S., and Namba, S. (1991). Program for calculating loudness according to din 45631 (iso 532b). *Journal of the Acoustical Society of Japan (E)*, 12(1) :39–42.

Table des figures

Liste des tableaux