Université de Bourgogne

Laboratoire d'Étude de l'Apprentissage et du Développement - CNRS UMR5022

École doctorale Environnements - Santé (ES)

THÈSE

En vue de l'obtention du grade de Docteure d'Université

Sous la direction de Maxime Ambard et Emmanuel Bigand

Développement d'un dispositif de substitution sensorielle vision-vers-audition : étude des performances de localisation et comparaison de schémas d'encodage

---

Présentée et soutenue publiquement par

Camille Bordeau

À Dijon, le 18 Décembre 2023, devant le jury composé de :

DR.	Daniel Pressnitzer	CNRS	*Président Rapporteur*
Dr.	Christian Graff	Université Grenoble Alpes	*Rapporteur*
DR.	Malika Auvray	CNRS	*Examinatrice*
Dr.	Cyrille Migniot	Université de Bourgogne	*Examinateur*
Dr.	Maxime Ambard	Université de Bourgogne	*Co-directeur de thèse*
Pr.	Emmanuel Bigand	Université de Bourgogne	*Directeur de thèse*

Résumé

Les dispositifs de substitution sensorielle vision-vers-audition convertissent des informations visuelles en un paysage sonore dans le but de permettre de percevoir l'environnement à travers la modalité auditive lorsque la modalité visuelle est altérée. Ils constituent une solution prometteuse pour améliorer l'autonomie des personnes déficientes visuelles lors de leurs déplacements pédestres. Ce travail de thèse avait pour objectif principal de déterminer et d'évaluer un schéma d'encodage pour la substitution sensorielle permettant la perception spatiale 3-dimensionnelle en proposant des protocoles de familiarisation et d'évaluation dans des environnements virtuels plus ou moins complexes. Le premier objectif était de déterminer si la reproduction d'indices acoustiques pour la perception spatiale auditive était plus efficace que l'utilisation d'autres indices acoustiques impliqués dans des interactions audio-visuelles. La première étude a mis en évidence que la modulation de la hauteur tonale dans le schéma d'encodage permettait de compenser en partie les limites perceptives de la spatialisation pour la dimension de l'élévation. La deuxième étude a mis en évidence que la modification de l'enveloppe sonore pouvait permettre de compenser la perception compressée de la distance. Le deuxième objectif de ce travail de thèse était de déterminer dans quelle mesure le schéma d'encodage utilisé préservait les capacités de perception spatiale dans un environnement complexe composé de plusieurs objets. La troisième étude a mis en évidence que les capacités de ségrégation d'une scène visuelle complexe à travers le paysage sonore associé dépendaient de la signature spectrale spécifique des objets la composant lorsque la modulation de la hauteur tonale est utilisée comme indice acoustique dans le schéma d'encodage. Les travaux de cette thèse ont des implications pratiques pour l'amélioration des dispositifs de substitution concernant, d'une part, la possibilité de compenser les limites perceptives spatiales avec des indices acoustiques non-spatiaux dans le schéma d'encodage, et d'une autre part, la nécessité de réduire le flux d'informations auditives pour préserver les capacités de ségrégation du paysage sonore. Les protocoles de familiarisation et d'évaluation en environnement virtuel ayant été développés de sorte à être adaptés à la population déficiente visuelle, les travaux de cette thèse soulignent le potentiel des environnements virtuels pour évaluer précisément les capacités d'utilisation de dispositifs de substitution dans un contexte contrôlé et sécurisé.

Mots-clés : substitution sensorielle, perception spatiale auditive, environnement virtuel, analyse de scènes auditives, localisation auditive, déficience visuelle

Abstract

Visual-to-auditory sensory substitution devices convert visual information into soundscapes for the purpose of allowing the perception of the environment with the auditory modality when the visual modality is impaired. They constitute a promising solution for improving the autonomy of visually impaired people when traveling on foot. The main objective of this thesis work was to determine an encoding scheme for sensory substitution allowing 3-dimensional spatial perception by proposing familiarization and evaluation protocols in virtual environments with different complexities. The first aim was to determine whether the reproduction of acoustic cues for auditory spatial perception was more effective than the use of acoustic cues involved in audio-visual interactions. The first study demonstrated that the modulation of pitch in the encoding scheme could partly compensate for the perceptual limits of spatialization for the dimension of elevation. The second study showed that the modification of the sound envelope could partly compensate for the compressed perception of distance. The second objective was to determine to what extent the determined encoding scheme preserved spatial perception abilities in a complex environment where several objects were present. The third study demonstrated that the segregation capabilities of a complex visual scene through the soundscape depend on the specific spectral signature of the objects composing it when pitch modulation is used as an acoustic cue in the encoding scheme. The work of this thesis has practical implications for the improvement of substitution devices concerning, on the one hand, the possibility of compensating spatial perceptual limits with non-spatial acoustic cues in the encoding scheme, and on the other hand, the need to reduce the amount of auditory information to preserve the segregation abilities of the soundscape. The familiarization and evaluation protocols in a virtual environment having been developed to be adapted to the visually impaired population, the work of this thesis highlights the potential of virtual environments to precisely evaluate the abilities to use sensory substitution devices in a secure context.

Keywords: sensory substitution, auditory spatial perception, virtual environment, auditory scene analysis, auditory localization, visual impairment

À celui qui est parti et celui qui est arrivé

À Papi Albert et à piojito Matias

Remerciements

Mes premiers remerciements s'adressent à mes superviseurs, **Maxime** et **Emmanuel**. Je tiens à vous remercier de m'avoir accordé votre confiance dans ce projet. Maxime, en particulier, je tiens à t'exprimer ma gratitude pour m'avoir ouvert les portes de la psychoacoustique et pour ton encadrement pendant ces trois années de thèse. Ta présence et tes corrections m'ont été d'une aide précieuse pendant les périodes de rédaction. Enfin, je n'imaginais pas ne pas te remercier pour ton humanité et pour ta bienveillance dont tu as régulièrement fait preuve au fil des mois. On sous-estime parfois l'impact positif de recevoir des félicitations et des encouragements, mais je peux affirmer aujourd'hui que j'ai eu de la chance de bénéficier de ces marques de soutien.

J'adresse mes remerciements à **Christian Graff** et **Daniel Pressnitzer** de me faire l'honneur d'évaluer mon travail de thèse en tant que rapporteurs, et à **Malika Auvray** et **Cyrille Migniot** de me faire l'honneur d'évaluer mon travail de thèse en tant qu'examinatrice et examinateur.

Je remercie les personnes avec qui j'ai eu l'opportunité de collaborer dans le cadre du projet 3DSG : **Alessandro**, **Cyrille**, **Florian**, **Julien**, **Mathilde**, et **Stéphane**. Votre enthousiasme, votre motivation, et votre constante bonne humeur, ont contribué à créer un environnement de travail stimulant. Vos remarques et suggestions pertinentes ont souligné l'importance de la collaboration avec des personnes issues de diverses disciplines pour enrichir notre compréhension du sujet, et élargir nos perspectives.

Je remercie la **Région Bourgogne-Franche-Comté**, le **Fond Européen de Développement Régional** et **l'Union National des Aveugles et Déficients Visuels** pour avoir participé au financement du projet 3DSG et des travaux de cette thèse.

Je tiens à remercier **François Cabestaing** et **Nicolas Grimault** d'avoir accepté de faire partie de mon Comité Scientifique de Suivi de Thèse et pour vos retours constructifs. Merci également à **Manuel Blouin** pour tes précieux conseils.

Mes prochains remerciements s'adressent à mes voisins et voisines de bureau : **Clémence**, **Guillaume**, **Joris** et **Thomas**. Entre nos échanges constructifs, nos moments de rires pour sortir la tête du guidon, et votre rationalité pour apaiser mes doutes, je vous remercie de m'avoir offert cet agréable environnement de travail. Par-dessus tout, je tiens à vous remercier pour votre accueil à mon arrivée, et pour votre bienveillance et votre soutien constants.

Merci à mes autres collègues de l'I3M, en particulier celles et ceux qui, à mon arrivée ou jusqu'à la fin, m'ont donné des conseils pour mener à bien cette thèse : **Alice**, **Jean**, **Manon**,

Marius et **Noé**. Vous avez contribué à un environnement de travail jovial et bienveillant, que ce soit autour d'un café ou d'une gamelle le midi.

Je remercie mes autres collègues du LEAD, en particulier **Claudia**, **Damien**, **Eleanor**, **Florian**, **Gaëtan**, **Iva**, **Julie F**, **Julie T**, **Laetitia**, **Laurie**, **Prany**, **Téo**, **Yannick**, **Williams**, et **Zahra**, avec qui j'ai eu l'opportunité de partager des moments autour d'une chocolatine ou un verre. Je vous suis reconnaissante d'avoir participé à créer un environnement de travail convivial.

Je tiens à remercier **Corinne** et **Sandrine** de toujours vous être rendues disponibles pour répondre attentivement et avec application à mes questions administratives. Plus largement, je remercie le laboratoire d'accueil **LEAD** qui a toujours mis un point d'honneur à être attentif aux besoins matériels et à y répondre, contribuant ainsi à un environnement de travail de qualité.

Je tiens à adresser mes remerciements à ma famille, en particulier à mes parents, **Nelly** et **Guy,** mon frère **Thomas** et **Natalia**. Il est moins difficile de croire en soi quand d'autres croient en nous, c'est pourquoi je vous remercie pour votre soutien indéfectible. Je vous suis reconnaissante d'avoir toujours eu confiance en mes choix durant mes études supérieures, quels qu'ils étaient, et de m'avoir fourni les ressources nécessaires pour les concrétiser.

Enfin, je tiens à remercier ma seconde famille : mes amis et amies, ou devrais-je dire, **les copaings**. Depuis Cadalen, Toulouse, Bordeaux, Madrid, Paris ou Caen, vous avez su m'apporter à distance les rires et le soutien dont j'avais besoin pour mener à bien cette thèse. Malgré la distance, les joies du hasard ont fait que nous étions ensemble lorsque j'ai appris l'acceptation de mon premier article en tant que première auteure. Je n'aurais pas pu rêver mieux que de trinquer à cette étape à vos côtés, dans les rues de Lisbonne, un shooter de Ginja à la main. Alors je tiens à remercier **Kakou**, **Roman**, **Séverin** et **Vincent**, et j'adresse des remerciements particuliers à **Anna**, **Charlotte**, **Cyrielle**, **Léa**, **Perrine**, et **Valentine** pour votre présence et vos encouragements. Je vous remercie de m'avoir redonné confiance lorsque celle-ci s'essoufflait.

Financement

Cette thèse a été co-financée par le Conseil Régional de Bourgogne-Franche-Comté (2020_0335), le Fond Européen de Développement Régional (FEDER) (BG0027904), et l'Union Nationale des Aveugles et Déficients Visuels (UNADEV). Elle a été réalisée au sein du Laboratoire d'Étude de l'Apprentissage et du Développement (CNRS UMR5022).

Liste des publications réalisées dans le cadre de la thèse

Publication publiée

Bordeau, C., Scalvini, F., Migniot, C., Dubois, J., & Ambard, M. (2023). Cross-modal correspondence enhances elevation localization in visual-to-auditory sensory substitution. *Frontiers in Psychology, 14*. https://doi.org/10.3389/fpsyg.2023.1079998

Publications à soumettre

Bordeau, C., Scalvini, F., Migniot, C., Dubois, J., & Ambard, M. (XXXX). Distance perception with a visual-to-auditory substitution device: compensating for the compression bias using sound envelope.

Bordeau, C., Scalvini, F., Migniot, C., Dubois, J., & Ambard, M. (XXXX). Localization abilities with a visual-to-auditory substitution device are modulated by the spatial arrangement of the scene.

Autres publications

Scalvini, F., **Bordeau, C.**, Ambard, M., Migniot, C., & Dubois, J. (2022). Low-Latency Human-Computer Auditory Interface Based on Real-Time Vision Analysis. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. https://doi.org/10.1109/icassp43922.2022.9747094

Scalvini, F., **Bordeau, C.**, Ambard, M., Migniot, C., Argon, S. & Dubois, J. (2022). Visual-auditory substitution device for indoor navigation based on fast visual marker detection, *16th International Conference on Signal Image Technology and Internet Based Systems (SITIS)*. https://doi.org/10.1109/SITIS57111.2022.00029

Scalvini, F., **Bordeau, C.**, Ambard, M., Migniot, C. & Dubois, J. (2024). Outdoor Navigation Assistive System Based on Robust and Real-Time Visual–Auditory Substitution Approach. *Sensors 24*(1). https://doi.org/10.3390/s24010166

Scalvini, F., **Bordeau, C.**, Ambard, M., Migniot, C., Vergnaud, M., & Dubois, J. (2023). uB-VisioGeoloc: An Image Sequences Dataset of Pedestrian Navigation Including Geolocalised-

Inertial Information and Spatial Sound Rendering of the Urban Environment's Obstacles. http://dx.doi.org/10.2139/ssrn.4521793 (**en cours de révision**)

Carlini, A., **Bordeau, C**., Ambard, M. (XXXX). Auditory localization: a comprehensive practical review (**soumis dans Heliyon**)

Liste des communications réalisées dans le cadre de la thèse

Communications affichées à des congrès internationaux

Bordeau, C., Scalvini, F., Migniot, C., Dubois, J. & Ambard, M. (2022). Distance perception of object using visual-to-auditory sensory substitution: comparison of conversion methods based on sound intensity and envelope modulation, *21st Auditory, Perception, Cognition and Action Meeting (APCAM 2022)*, Boston, Massachusetts, Etats-Unis, Novembre 2022.

Bordeau, C., Scalvini, F., Migniot, C., Dubois, J. & Ambard, M. (2023). Visual-to-auditory conversion methods for sensory substitution: sound spatialization only versus cross-modal correspondence, *International Multisensory Research Forum (IMRF 2023)*, Bruxelles, Belgique, Juin 2023.

Bordeau, C., Scalvini, F., Migniot, C., Dubois, J. & Ambard, M. (2023). Investigation of the effect of distractors on localization abilities with a visual-to-auditory substitution device, *International Multisensory Research Forum (IMRF 2023)*, Bruxelles, Belgique, Juin 2023.

Communication orale à un congrès national

Bordeau, C., Scalvini, F., Migniot, C., Dubois, J. & Ambard, M. (2022). Comparison of visual-to-auditory conversion methods for sensory substitution in a localization task. *Journée des doctorants du Laboratoire d'Etude de l'Apprentissage et du Développement*, Dijon, France, Juin 2022.

Communication affichée à un congrès national

Bordeau, C., Scalvini, F., Migniot, C., Dubois, J. & Ambard, M. (2022). Comparison of auditory encodings for visual-to-auditory sensory substitution in a localization task, *Forum des Jeunes Chercheurs de l'Ecole Doctorale Environnement-Santé*, Dijon, France, Juin 2022.

Communications de vulgarisation scientifique

Bordeau, C. Des lunettes pour aider les personnes aveugles à localiser des obstacles grâce à des sons. *Experimentarium*, Dijon, France, Mars 2022 & Mars 2023. https://www.experimentarium.fr/les-chercheurs/des-lunettes-pour-aider-les-personnes-aveugles-localiser-des-obstacles-grace-des-sons. Initié par l'Université de Bourgogne, l'Experimentarium est un programme de rencontres entre jeunes chercheurs et chercheuses, le grand public, et le jeune public à travers des ateliers d'échanges et des conférences.

DES LUNETTES POUR AIDER LES PERSONNES AVEUGLES À LOCALISER DES OBSTACLES GRÂCE À DES SONS

→ PSYCHOLOGIE COGNITIVE & ACOUSTIQUE

CAMILLE BORDEAU est jeune chercheuse en Psychologie Cognitive et Acoustique au LEAD*, à Dijon. Avec plusieurs chercheurs de son équipe, elle participe à l'invention de lunettes particulières pour les personnes aveugles, qui leur permettraient de détecter des obstacles. Ces lunettes ne contiennent pas de verres, mais une caméra et des écouteurs fixés sur la monture. Elles convertissent ce qui se trouve devant la personne en plusieurs sons. Le rôle de Camille est de trouver quels sons choisir pour que les personnes localisent facilement les obstacles.

* Laboratoire d'Etude de l'Apprentissage et du Développement

« Quand je croise une personne aveugle dans la rue, je me dis qu'un jour, peut-être, elle écoutera aussi les sons que j'ai créés et que j'écoute toute la journée ! Mais elle, ce sera pour l'aider à se déplacer en sécurité. »

Camille Bordeau

L'EXPÉ www.experimentarium.fr

En France, il y a plus de 200 000 personnes aveugles. Leurs problèmes de vision ont un impact sur leur vie de tous les jours. Par exemple, lorsqu'elles se déplacent, c'est plus difficile pour elles de savoir où sont les obstacles, et cela peut être dangereux.

À Dijon, une dizaine de chercheurs sont en train d'inventer des lunettes spéciales pour aider les personnes aveugles à détecter des obstacles en les alertant grâce à des sons. C'est un peu comme si on remplaçait la vision par l'audition. On appelle ça la « substitution sensorielle ».

Camille participe à ce projet. Elle étudie comment informer au mieux l'endroit où se situe un obstacle. Quels sons choisir pour indiquer si un obstacle est en haut ou en bas, à gauche ou à droite, proche ou loin ? C'est ce que cherche Camille en créant plusieurs séries de sons possibles, plus ou moins compliquées.

Pour tester ces sons, Camille a créé une sorte de jeu vidéo dans lequel elle place de faux obstacles. Elle fait venir des participants pour jouer à ce jeu les yeux fermés. Elle place les faux obstacles dans le jeu et elle leur demande de les localiser en utilisant les différents types de son qu'elle a créés. Camille regarde ensuite si les participants ont mieux réussi à repérer où étaient les obstacles avec des sons simples ou compliqués.

Ensuite, Camille espère pouvoir tester ces lunettes auprès de personnes aveugles, avec des vrais obstacles dans un pièce puis dehors, dans la rue.

LES OBJECTIFS

- Inventer des lunettes particulières pour aider les personnes aveugles à détecter des obstacles grâce à des sons
- Identifier si c'est plus facile de localiser des obstacles avec des sons simples ou compliqués
- Créer une sorte de jeu vidéo qui se joue les yeux fermés pour tester les lunettes dans des situations sans danger

L'EXPÉ www.experimentarium.fr

Bordeau, C., Scalvini, F., Ambard, M. Ouvrir les oreilles pour mieux voir. *5ème journée de culture scientifique à l'Institut Marey*, Dijon, France, 28 mars & 2 mai 2023

Table des matières

Liste des figures

Liste des tableaux

Liste des abréviations

DSS : Dispositif de substitution sensorielle (ou SSD, *Sensory substitution device* en anglais)

HRTF : Fonction de transfert relative à la tête (pour *Head-related transfer function* en anglais)

ILD : Différence interaurale en intensité (pour *Interaural level difference* en anglais)

ITD : Différence intéraurale temporelle (pour *Interaural time difference* en anglaise)

IPD : Différence interaurale de phase (pour *Interaural phase difference* en anglais)

SNR : Rapport signal-sur-bruit (pour *Signal-to-noise ratio* en anglais)

I. Introduction générale

La déficience visuelle a des répercussions considérables sur l'inclusion sociale et la qualité de vie des personnes touchées. L'Organisation Mondiale de la Santé distingue plusieurs catégories de déficience visuelle définissant le degré d'atteinte : la déficience légère, moyenne, sévère, et la cécité visuelle (World Health Organization, 2019). La déficience visuelle légère concerne les personnes ayant une acuité visuelle entre 6/12 et 20/60 et un champ visuel supérieur à 10°. La déficience visuelle moyenne concerne les personnes ayant une acuité visuelle entre 6/60 et 20/60, et un champ visuel supérieur à 10°. La déficience visuelle sévère concerne les personnes ayant une acuité visuelle entre 3/60 et 6/60, et un champ visuel supérieur à 10°. Enfin, la cécité visuelle concerne les personnes ayant un champ visuel inférieur à 10° et/ou une acuité visuelle inférieure à 3/60, avec ou sans perception de la lumière. En 2020, on estimait à au moins 337 millions dans le monde le nombre de personnes déficientes visuelles moyennes, sévères et en cécité visuelle (Steinmetz et al., 2021), dont 43 millions étaient atteintes de déficience visuelle sévère ou de cécité. Les principales causes de la déficience visuelle sont la cataracte, le glaucome, les erreurs de réfraction, la dégénérescence maculaire liée à l'âge et la rétinopathie diabétique (Steinmetz et al., 2021 ; World Health Organization, 2019).

Avec l'accroissement de la population mondiale accompagné de l'augmentation de l'âge moyen, le nombre de personnes atteintes de déficience visuelle et de cécité est malheureusement en progression (Steinmetz et al., 2021 ; World Health Organization, 2019). Différentes actions sont mises en place au niveau international : certaines sont préventives, d'autres thérapeutiques (développement de traitements et amélioration de l'accessibilité à ces traitements), et d'autres concernent la réhabilitation. La déficience visuelle a de graves répercussions sociales, financières, physiques et psychologiques sur le quotidien des personnes touchées. Elle entraîne une augmentation du risque de chutes et de blessures, une diminution de l'accès au monde du travail et aux loisirs, et de l'autonomie dans les tâches quotidiennes (e.g. faire les courses, la cuisine) et les déplacements (Burton et al., 2021 ; World Health Organization, 2019). Ces éléments participent à un risque accru d'isolement social et à la diminution de l'autonomie et de la qualité de vie.

Améliorer l'autonomie des personnes non-voyantes lors de leurs déplacements représente un enjeu majeur pour améliorer leur qualité de vie. Malgré les dispositifs actuels d'aide à la locomotion pour guider et garantir la sécurité dans les déplacements (e.g., la canne blanche, le chien d'aveugle, les GPS sur smartphone), les problèmes persistent. Une autre catégorie de dispositifs d'assistance pour les personnes non-voyantes se révèle prometteuse : les dispositifs de substitution sensorielle (DSSs). Ces dispositifs convertissent des informations visuelles en informations dans une modalité alternative de substitution (auditive ou tactile). Conceptuellement, le principe de conversion peut être comparé à celui du Braille, qui permet de « lire avec les doigts » en

convertissant des informations visuelles (i.e., lettres et symboles) en informations haptiques perçues grâce au déplacement des doigts sur des points en relief. Dans les années 1960, Bach-y-Rita développe un dispositif innovant qui convertit des informations visuelles acquises avec une caméra sous la forme de vibrations tactiles (Bach-Y-Rita et al., 1969). Il s'agit du premier dispositif de substitution sensorielle vision-vers-tactile. Bach-y-Rita démontre alors la possibilité de percevoir des informations initialement visuelles par le biais d'une modalité sensorielle alternative dite de substitution. Une vingtaine d'années plus tard, le premier DSS vision-vers-audition, the vOICe (Meijer, 1992) est développé. Avec ce dispositif, les informations visuelles, sous forme d'un flux vidéo acquis avec une caméra, sont converties en un paysage sonore en suivant un schéma d'encodage mettant en correspondance des dimensions visuelles (position horizontale et verticale, et luminosité) avec des indices acoustiques (scanning temporel, fréquence et intensité). Depuis sa création, le potentiel de the vOICe pour effectuer des tâches de reconnaissance d'objets, de localisation d'objets ou de navigation a été démontré chez des personnes voyantes et non-voyantes. Entre temps, d'autres dispositifs de substitution vision-vers-audition ont été développés, utilisant une variété d'indices acoustiques dans le schéma d'encodage vision-vers-audition.

Malgré le nombre grandissant de dispositifs de substitution sensorielle vision-vers-audition, ils ne sont que très peu adoptés par la population non-voyante ciblée (Elli et al., 2014). Parmi les raisons expliquant ce manque d'engouement, on retrouve le coût onéreux des dispositifs et la nécessité de porter du matériel encombrant. Au-delà de ces raisons économiques et pratiques, c'est majoritairement la question de l'utilisabilité des dispositifs qui est pointée du doigt. Maîtriser l'utilisation d'un tel dispositif implique d'apprendre à interpréter de nouvelles informations sensorielles auditives pour percevoir le monde et interagir avec. En ce sens, il s'agit d'un apprentissage perceptif (Ahissar et al., 2009) qui nécessite l'intégration de contingences sensorimotrices (Briscoe, 2018). La mise en correspondance entre les dimensions visuelles et les indices acoustiques modulés dans les schémas d'encodage n'est pas toujours considérée comme intuitive par la population cible (Hamilton-Fletcher et al., 2016b). Conséquemment, les entraînements, souvent longs, peuvent être à l'origine d'un découragement freinant l'adoption du dispositif. À terme, il s'agit d'être en capacité d'analyser des scènes auditives complexes (Bregman, 1990), puisqu'il est nécessaire de percevoir les informations auditives provenant à la fois du paysage sonore du dispositif, et de l'environnement réel (e.g., bruit de voitures), sans qu'elles entrent en conflit. Ces éléments démontrent la place centrale du choix des indices acoustiques à utiliser dans le schéma d'encodage du dispositif, et la nécessité de prendre en compte les capacités d'analyse de scènes auditives.

C'est en partant de ces constats que le projet 3D Sound Glasses (3DSG) a été initié. Ce projet, financé par le Conseil Régional de Bourgogne-Franche-Comté (2020_0335), le Fond Européen de Développement Régional, FEDER (BG0027904) et l'Union Nationale des Aveugles et Déficients Visuels (UNADEV), vise à développer un dispositif de substitution sensorielle vision-vers-audition dont la fonctionnalité est l'aide à la locomotion des personnes non-voyantes. Le projet repose sur une étroite collaboration interdisciplinaire entre un laboratoire en Informatique (ImViA, Image et Vision Artificielle) et un laboratoire en Psychologie (LEAD, Laboratoire d'Étude de l'Apprentissage et du Développement). Les travaux au sein de l'ImViA se concentrent sur la conception du dispositif d'un point de vue implémentation algorithmique de traitement vidéo en temps réel, ainsi que sur l'implémentation électronique du dispositif qui doit être utilisable pendant des déplacements. Au sein du LEAD, les travaux s'intéressent aux contraintes cognitives et perceptives dans le développement du dispositif de substitution. En s'intégrant au projet 3DSG, la présente thèse en Psychologie vise spécifiquement à optimiser les informations auditives transmises par le dispositif de substitution pour qu'elles soient à la fois en adéquation avec les capacités humaines de perception, et pertinentes pour une fonctionnalité d'aide à la locomotion et à la localisation d'obstacles. **Ce travail de thèse a pour objectif d'évaluer des schémas d'encodage pour la substitution sensorielle vision-vers-audition permettant la perception spatiale, en proposant des protocoles d'évaluation pour comparer plusieurs schémas d'encodage dans des environnements virtuels plus ou moins complexes**. Les travaux visent d'une part à proposer des schémas d'encodage adaptés aux capacités humaines de perception auditive et de perception spatiale afin de faciliter l'apprentissage perceptif permettant l'analyse de scènes auditives complexes, et d'une autre part, à proposer des protocoles de familiarisation facilitant l'acquisition de contingences sensorimotrices.

Le présent manuscrit s'articule autour de trois parties. La première partie présentera le cadre théorique de ce travail. La nécessité de développer les dispositifs de substitution sensorielle vision-vers-audition au regard des capacités humaines de perception auditive spatiale sera abordée, et ces capacités seront présentées. Les capacités de perception spatiale avec les dispositifs existants seront finalement présentées. L'hétérogénéité des dispositifs existants et des méthodes d'évaluation sera abordée, soulevant la pertinence de développer des protocoles réplicables en réalité virtuelle pour évaluer et comparer les capacités de perception avec ces dispositifs. La deuxième partie du manuscrit présentera les travaux de recherche réalisés au cours de la thèse, ce qui comprend trois études investiguant les capacités de localisation d'objets virtuels en fonction du schéma d'encodage du dispositif et de la complexité de l'environnement virtuel. Enfin, la dernière partie sera consacrée à la conclusion générale de ce travail de thèse. Elle discutera les résultats obtenus, leur implication pour le développement de dispositifs de substitution, et des perspectives qu'ils ouvrent.

II. Cadre théorique

1. La substitution sensorielle vision-vers-audition

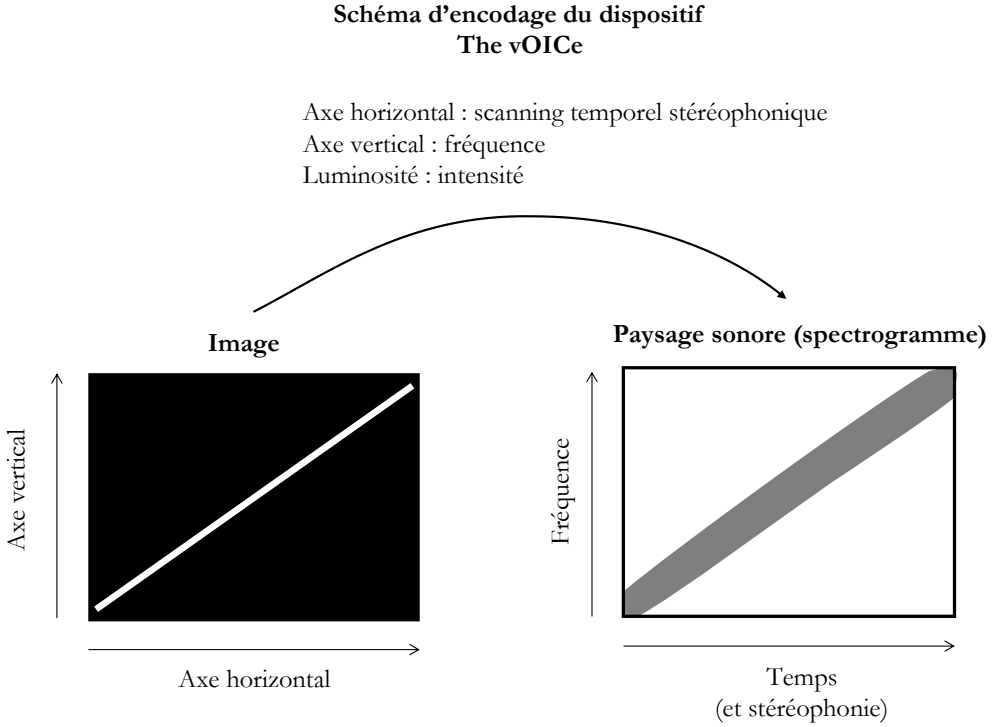
1.1. Définitions et principes

La substitution sensorielle consiste à utiliser une modalité sensorielle alternative (la modalité de substitution) pour percevoir des informations issues d'une modalité sensorielle d'origine (la modalité à substituer). Les dispositifs de substitution sensorielle (DSSs) ont initialement été conçus pour assister les personnes non-voyantes et malvoyantes en les aidant à percevoir leur environnement. Dans ce cadre, la modalité sensorielle à substituer est bien entendu la vision. Parmi les dispositifs existants, les modalités tactile ou auditive sont utilisées comme modalités de substitution. Ainsi, le principe des DSSs repose sur la conversion d'images en stimulations tactiles ou en paysages sonores mettant en correspondance les caractéristiques visuelles d'une image (e.g. localisation, couleur, luminosité…) avec des patterns tactiles (e.g. fréquence de vibrations, amplitude, localisation) ou des indices acoustiques (e.g. fréquence de sons, volume, timbre, localisation).

Historiquement, le premier dispositif développé est un DSS vision-vers-tactile, appelé TVSS (Tactile Vision Sensory Substitution) qui a été proposé par Bach-Y-Rita et al. (1969). Avec le TVSS, les images, acquises avec une caméra vidéo, sont converties en stimulations vibro-tactiles délivrées sur le dos. À l'issue d'une période d'entraînement pour apprendre à utiliser le dispositif, Bach-y-Rita rapporte que les participants ne percevaient désormais plus les sensations comme étant proximales, mais comme provenant de l'environnement externe. Bach-y-Rita propose alors que le TVSS permet de « voir » avec la modalité tactile.

Une vingtaine d'années plus tard, Meijer (1992) propose un DSS qui convertit des images en sons : le dispositif the vOICe. Il s'agit du premier DSS vision-vers-audition. The vOICe convertit une image en un paysage sonore en suivant un schéma d'encodage modulant des indices acoustiques telles que la fréquence, l'intensité et la temporalité (**Figure II-1**). Le paysage sonore prend la forme d'un son complexe d'une durée d'une seconde, composé d'une combinaison de plusieurs tonalités pures dont l'intensité varie en fonction de la luminosité des pixels de l'image. Dans son schéma d'origine, the vOICe transmet donc auditivement l'image en la scannant de gauche à droite, en encodant la position verticale avec la fréquence du son, la position horizontale par le *scanning* temporel et la stéréophonie, et la luminosité par l'intensité. Ainsi, chaque ligne de pixels de l'image est associée à une fréquence, allant de 500 à 5000 Hz.

Figure II-1. Représentation du schéma d'encodage utilisé dans le dispositif de substitution sensorielle vision-vers-audition the vOICe (Meijer, 1992). Il convertit 3 dimensions de l'image (axes horizontal et vertical, et la luminosité) en indices acoustiques. Le paysage sonore associé à une image sur laquelle figure une diagonale ascendante prend la forme d'un stimulus auditif d'une durée de 1 seconde dont la fréquence augmente au cours de la diffusion pour passer de 500 à 5000 Hz.



L'utilisation de la modalité auditive comme modalité de substitution présente de nombreux avantages car le système auditif permet une perception fine, et la multidimensionnalité du son permet de combiner la modulation de plusieurs indices acoustiques (D. M. Howard & Angus, 2009). Si la modalité auditive est parfois combinée avec la modalité tactile dans un DSS, la modalité auditive a été qualifiée de plus intuitive dans le cadre d'un tel dispositif (Lupu et al., 2020).

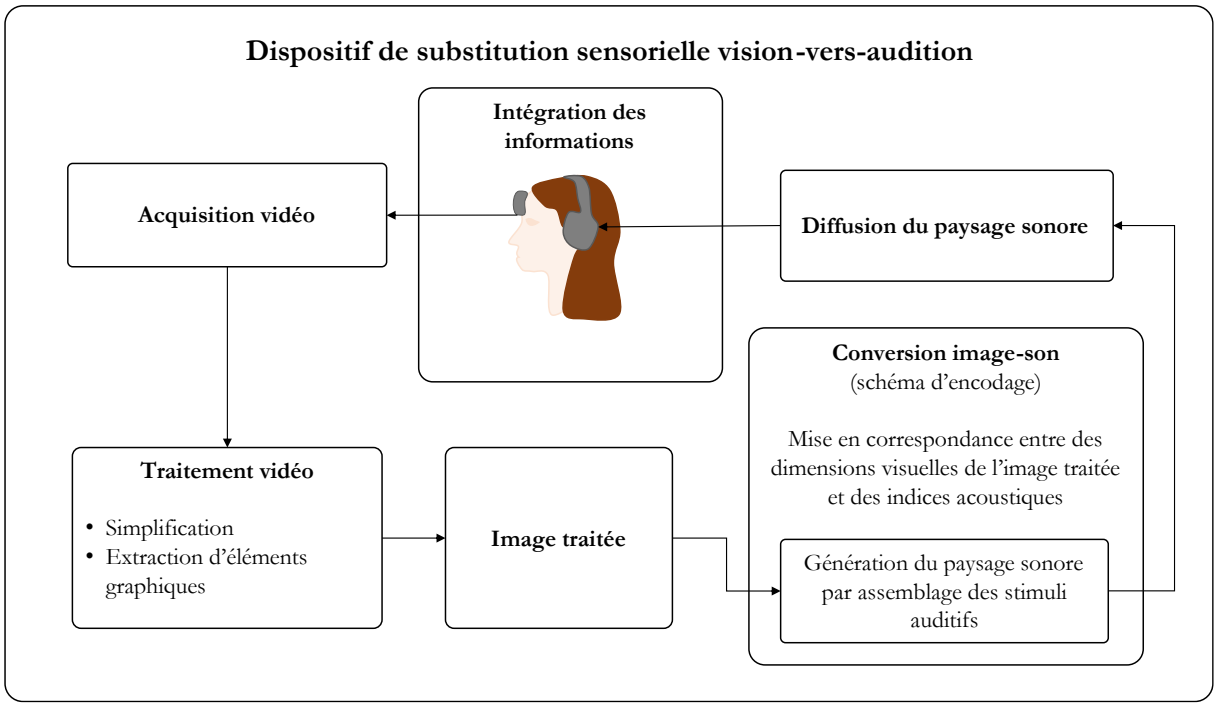
Depuis le premier DSS the vOICe il y a une trentaine d'années, de nombreux dispositifs de substitution sensorielle vision-vers-audition ont été développés mais reposant tous sur un même principe : celui de la conversion d'images acquises par une caméra vidéo en un paysage sonore. Le fonctionnement général d'un DSS vision-vers-audition est schématisé sur la **Figure II-2**.

La première étape de ces dispositifs consiste en un traitement du flux vidéo acquis par la caméra. Le traitement de l'image a pour objectif de simplifier les informations visuelles (e.g., en redimensionnant l'image), et d'extraire les éléments graphiques d'intérêt (e.g., couleur, contours d'objets, emplacements d'objets). L'image traitée est ensuite convertie dans une seconde étape en un paysage sonore en suivant un schéma d'encodage qui associe des éléments graphiques et des dimensions visuelles à des indices acoustiques. Les indices acoustiques sont ainsi modulés en fonction des caractéristiques graphiques de l'image traitée. Bien souvent, chaque pixel de l'image traitée est associé à un signal acoustique caractérisant son emplacement (i.e., dimension spatiale

visuelle) et un élément graphique autre (e.g., couleur, luminosité). Les éléments graphiques transmis et les indices acoustiques modulés dans les DSSs existants sont variés, et sont détaillés dans la section II.3.1.

Sans la vision, les personnes non-voyantes rencontrent des difficultés pour percevoir et interagir avec l'environnement autour d'eux. L'objectif des DSSs est de permettre à des fonctionnalités normalement réalisées grâce à la modalité visuelle d'être assurées par une autre modalité sensorielle, pour ainsi pallier la privation visuelle. Ces fonctionnalités sont variées telles que la reconnaissance d'objets, leur localisation, leur saisie, ou encore l'évitement d'obstacles. L'idée est de tendre vers une équivalence fonctionnelle entre les deux modalités sensorielles, et le développement des DSSs se fait généralement au regard de ces besoins fonctionnels, bien que la restauration complète d'une fonctionnalité soit bien évidemment inatteignable (Auvray et al., 2019). Les études conduites jusqu'à présent mettent en évidence la multiplicité des tâches qui peuvent être effectuées avec un DSS. Parmi ces tâches, on retrouve des tâches de reconnaissance d'objets réels (Auvray et al., 2007 ; Brown et al., 2011), de formes (Abboud et al., 2014 ; Arno et al., 2001 ; Bermejo et al., 2015 ; Brown et al., 2015), de couleurs (Abboud et al., 2014 ; Levy-Tzedek et al., 2012), de navigation (Jicol et al., 2020 ; Neugebauer et al., 2020 ; Paré et al., 2021 ; Pasqualotto & Esenkaya, 2016), et de localisation d'objets (Ambard et al., 2015 ; Auvray et al., 2007 ; Levy-Tzedek et al., 2012 ; Renier & De Volder, 2010 ; Richardson et al., 2019).

Figure II-2. Schématisation du fonctionnement d'un DSS vision-vers-audition en temps réel. La caméra portée par l'utilisateur acquiert un flux vidéo qui est ensuite traitée. L'image traitée simplifiée est ensuite convertie en un paysage sonore en suivant un schéma d'encodage. Le paysage sonore est diffusé à l'utilisateur qui intègre ces informations auditives.



1.2. Percevoir avec un dispositif de substitution : questions phénoménologiques et intégration verticale

Effectuer une tâche avec un DSS vision-vers-audition nécessite de pouvoir interpréter les sensations dans la modalité alternative qu'est la modalité auditive. D'un point de vue phénoménologique, les perceptions issues de ces sensations peuvent être qualifiées différemment. Considérons par exemple une tâche de localisation d'objets avec un DSS vision-vers-audition. Les sensations provenant du dispositif sont certes auditives : le paysage sonore transmis induit des sensations propres à la modalité auditive (i.e., l'énergie contenue dans le signal acoustique pour différentes fréquences, la durée de stimulation…). Mais qu'en est-il de la perception qui en résulte, c'est-à-dire de la représentation mentale issue de ces sensations. Est-elle purement auditive ? Assimilable à la modalité visuelle ? Issue d'une intégration supra-modale ?

Plusieurs hypothèses ont été évoquées concernant la modalité sensorielle à laquelle ces perceptions s'apparentent (e.g Auvray & Myin, 2009 ; Briscoe, 2018 pour revue). Initialement, deux approches s'opposaient pour qualifier la perception dans le contexte de la substitution sensorielle : la thèse de la déférence et la thèse de la dominance. Alors que l'approche par la déférence stipule que la perception avec un DSS s'apparente à une perception dans la modalité substituée (i.e., modalité visuelle, e.g. Hurley & Noë, 2003), la thèse de la dominance prône que la perception avec un DSS se restreint à une perception dans la modalité de substitution (i.e., auditive dans notre cas, e.g. Block, 2003). La thèse de la dominance repose sur le fait que la perception repose essentiellement sur les sensations, affirmant la place centrale de la modalité auditive. Néanmoins, une modalité sensorielle ne se définit pas uniquement sur le critère des sensations, mais aussi vis-à-vis de l'organe sensoriel impliqué, des propriétés physiques des stimuli, de l'expérience qualitative, et de l'équivalence comportementale et sensorimotrice (Auvray & Myin, 2009). À travers ces critères, plusieurs éléments suggèrent que la perception avec un DSS peut s'apparenter à une perception visuelle, telle que l'équivalence sensori-motrice et comportementale (Bermejo et al., 2015 ; O'Regan & Noë, 2001), ou bien l'activation d'aires dites visuelles lors de l'utilisation de DSSs (Kupers et al., 2006 ; Renier et al., 2005). L'équivalence entre la perception avec un DSS et à travers la modalité visuelle (ou inversement, auditive) n'étant pas directe, la théorie de l'intégration verticale nuance ces deux approches en stipulant que la perception avec un DSS doit être considérée comme une nouvelle forme de perception dans le sens où elle s'apparente à une extension de perception (Arnold et al., 2017 ; Auvray, 2019 ; Auvray & Myin, 2009).

La théorie de l'intégration verticale se place dans la lignée de la théorie du cerveau métamodal (Pascual-Leone & Hamilton, 2001), qui stipule que l'intégration des informations sensorielles serait davantage fonction-dépendante, c'est-à-dire dépendante de la tâche effectuée

plutôt que dépendante de la modalité sensorielle impliquée. Ainsi, Auvray et Myin (2009) considèrent que la perception avec un DSS s'apparenterait davantage à une extension fonctionnelle, permettant d'effectuer une tâche en se reposant sur des informations sensorielles d'une nouvelle modalité. Avec l'approche de la théorie de l'intégration verticale, la perception à travers un DSS dépend à la fois des capacités de perception dans la modalité de substitution (i.e., auditive) mais également des capacités propres à la modalité substituée (e.g., expérience visuelle passée). Ainsi, la perception avec un DSS reposerait sur des capacités préexistantes cognitives et de perception dans les deux modalités, qui permettraient de construire une nouvelle façon de percevoir en permettant l'extraction des percepts à partir de ces nouvelles sensations auditives (Auvray, 2019).

La théorie de l'intégration verticale prédit ainsi des variabilités interindividuelles en termes de mécanismes cognitifs impliqués dans la perception avec un DSS puisque la perception avec un tel dispositif reposerait sur des capacités préexistantes impliquant des processus cognitifs spécifiques à une des deux modalités sensorielles et non-spécifiques (supramodales) (Arnold et al., 2017). Cette approche est appuyée par des études mettant en évidence des corrélations entre des capacités préexistantes propres à la modalité de substitution (i.e., auditive) ou propres à la tâche (e.g., perception spatiale, mémoire de travail…) avec les capacités à utiliser un DSS. Par exemple, plusieurs études ont suggéré que les capacités auditives de bas niveau pouvaient influencer les performances dans des tâches impliquant un DSS vision-vers-audition, améliorant les performances en accélérant l'apprentissage. Par exemple, dans Pesnot Lerousseau et al. (2021), les capacités de discrimination de 3 dimensions du son (l'intensité, la fréquence, la durée) et d'extraction de sons complexes parmi un fond sonore bruité étaient évaluées. Ces capacités de discrimination et d'extractions étaient corrélées avec une meilleure utilisation du DSS. Dans Haigh et al. (2013), la pratique d'un instrument de musique par le passé était corrélée à de meilleures performances, mais le score de discrimination de hauteur tonale ne l'était pas. Également, lors d'un apprentissage perceptif impliquant l'interprétation de nouvelles informations auditives provenant de la conversion d'informations visuelles, Hanneton et al. (2015) ont mis en évidence une corrélation positive entre le score à un test de rotation mentale visuelle et les performances dans la tâche nécessitant l'interprétation de ces nouvelles informations auditives. Pourtant, ces corrélations entre les capacités préexistantes et les performances d'utilisation d'un DSS (ou d'un dispositif s'y apparentant) ne semblent pas systématiques. Par exemple, Brown et al. (2011) et Haigh et al. (2013) ne montrent pas de lien entre les capacités d'imagerie visuelle et les performances à une tâche impliquant le dispositif The vOICe.

Ainsi, les variabilités interindividuelles dans les performances d'utilisation d'un DSS en lien avec les capacités préexistantes soulignent l'importance de la prise en compte des capacités perceptives humaines et de leur hétérogénéité lors du développement d'un DSS.

1.3. Apprendre à utiliser un dispositif de substitution : un apprentissage perceptif sensorimoteur en 5 étapes

L'appropriation d'un DSS repose sur la capacité à extraire et interpréter les nouvelles informations sensorielles auditives issues du DSS. L'apprentissage perceptif consiste en l'amélioration des performances dans des tâches perceptuelles, et il repose sur la pratique de la tâche (Ahissar et Hochstein, 2004). En ce sens, l'appropriation d'un DSS est un apprentissage perceptif durant lequel la perception va spécifiquement s'affiner pour l'utilisation du DSS. D'après la théorie de la hiérarchie inverse pour l'apprentissage perceptif (Ahissar et al., 2009), le traitement des informations sensorielles se fait à plusieurs niveaux de traitement qui sont hiérarchisés, allant de bas niveaux à plus hauts niveaux. Les représentations de hauts niveaux sont celles porteuses de sens, et émergent à partir de l'association entre plusieurs représentations de bas niveaux telles que des informations spectro-temporelles fines pour la modalité auditive. Dans cette théorie, la perception passe principalement par les représentations de haut niveau qui sont adaptées écologiquement, mais les relations entre les différents niveaux de traitement sont modulables et adaptables, rendant possible l'apprentissage perceptif. Dans le contexte de la substitution sensorielle vision-vers-audition et selon cette théorie, l'apprentissage pourrait consister à sélectionner les indices acoustiques de bas niveaux pertinents pour la tâche en question (e.g., tâche de localisation) et à affiner la pondération de ces représentations de bas niveaux menant aux représentations de hauts niveaux concernées. En se reposant sur des capacités cognitives et perceptives préexistantes, il s'agit, à terme, de pouvoir accéder à des représentations spatiales et cognitives de haut niveau avec le dispositif (Auvray, 2019). Dans le contexte d'un DSS visant une fonctionnalité d'aide à la locomotion, l'apprentissage repose sur une recalibration des indices acoustiques utilisés dans le DSS par rapport à une représentation spatiale préexistante (Nardini, 2021).

L'appropriation d'un DSS s'apparente à un processus d'externalisation qui peut être décrit en cinq étapes (Auvray, 2004) : le contact, l'attribution distale, la maîtrise de l'espace distal, la localisation distale, et la constitution d'une expérience distale. Durant la première étape (le contact), l'utilisateur extrait les régularités sensorimotrices et cette étape est facilitée par l'action motrice active (Auvray et al., 2005). La deuxième étape (l'attribution distale) consiste à la prise de conscience que les changements dans les sensations auditives perçues proviennent d'un objet distant dans l'espace. L'utilisateur peut alors apprendre à maîtriser cet espace distal (troisième étape) en

développant sa capacité à faire varier son point de vue distal. Ensuite, l'étape de la localisation distale (quatrième étape) consiste en une autonomisation de l'attribution distale. Enfin, la dernière et cinquième étape, celle de la constitution d'une expérience distale donne lieu à la création de sens vis-à-vis de cette nouvelle perception.

L'apprentissage de l'utilisation d'un DSS implique donc d'apprendre à associer les sensations auditives provenant du paysage sonore et ses modifications en fonction de ses propres actions motrices. En ce sens, il s'agit d'un apprentissage sensorimoteur reposant sur l'association entre des actions motrices et des modifications au niveau des sensations auditives. À terme, les sensations perçues à travers le DSS sont attribuées à des éléments extérieurs qui vont au-delà de la stimulation sensorielle auditive qui est proximale, i.e. l'attribution distale (Auvray, 2004 ; Briscoe, 2018). Elle repose notamment sur l'extraction de contingences sensorimotrices (Chebat et al., 2017), qui sont des régularités mettant en lien les actions motrices avec des changements concernant une stimulation sensorielle proximale (Briscoe, 2018 ; Chebat et al., 2017). L'exploration active et volontaire de l'environnement par des actions motrices avec le DSS est alors centrale dans l'acquisition de ses contingences sensorimotrices, permettant ainsi de fermer la boucle sensorimotrice. Cette nécessité d'une action motrice a été relevée dès le premier dispositif développé, le DSS vision-vers-tactile TVSS (Bach-Y-Rita et al., 1969). Dans White et al. (1970), seuls les participants ayant la possibilité de contrôler activement la caméra (à l'inverse d'une exposition passive aux stimulations tactiles) performaient à la tâche avec le TVSS. Ultérieurement, en comparant des actions exploratoires auto-générées et des actions exploratoires avec des stimulations d'un DSS vision-vers-tactile pré-enregistrées, Diaz (2012) a mis en évidence que l'action motrice permettait l'intégration du lien entre les mouvements de la caméra et les sensations perçues. Avec le DSS vision-vers-audition the Vibe, Auvray et al. (2005) ont également mis en évidence les bénéfices d'un couplage entre ses propres mouvements auto-générés et les conséquences sur les sensations auditives provenant du DSS. Selon les auteurs, c'est l'intégration de ces contingences sensorimotrices qui permet l'émergence de la perception distale (Auvray, 2004 ; Auvray et al., 2005). Le contrôle actif de la caméra par laquelle les images sont acquises serait essentiel pour l'attribution distale car il permettrait à la fois l'adoption d'un point de vue égocentrique, et l'élaboration d'une représentation interne de l'emplacement des objets à distance et de la façon dont ils sont arrangés spatialement (Briscoe, 2018).

Ainsi, apprendre à utiliser un DSS semble nécessiter un apprentissage perceptif sensorimoteur facilité, voir conditionné, par la présence d'actions motrices permettant d'extraire les contingences sensorimotrices nécessaires à l'intégration des nouvelles informations sensorielles et à l'attribution distale.

1.4. Freins à l'adoption des dispositifs de substitution et préconisations pour leur développement

Malgré la multiplicité des tâches pouvant être effectuées avec un DSS, la faisabilité des DSSs démontrée en laboratoire ne se traduit pas par une large adoption des DSSs par la population non-voyante (Elli et al., 2014 ; Kristjánsson et al., 2016 ; Lloyd-Esenkaya et al., 2020 ; Maidenbaum et al., 2014a). Plusieurs raisons sont évoquées pour expliquer le peu de personnes non-voyantes utilisant quotidiennement un DSS.

Jusqu'à présent, les DSSs sont majoritairement étudiés en laboratoire, dans des conditions contrôlées laissant peu de place à une projection de leur utilisabilité dans un environnement réel plus complexe. Par exemple, être capable de reconnaître avec un DSS une forme telle qu'une ligne diagonale blanche sur fond noir présentée sur un écran (e.g., **Figure II-1**) peut difficilement être généralisable à la capacité de se représenter mentalement une scène urbaine complexe. Pour répondre à ce frein, l'évaluation des DSSs dans des conditions plus écologiques telles qu'une rue semble nécessaire, mais un environnement urbain implique des risques tels que des collisions et rend difficile le contrôle des variables. En ce sens, les environnements virtuels et augmentés offrent des possibilités intéressantes (Maidenbaum et al., 2014a), en permettant de manipuler la complexité de l'environnement tout en garantissant la sécurité et la reproductibilité. À la différence des études utilisant des stimuli sur écrans, les environnements virtuels offrent la possibilité de fermer la boucle sensori-motrice, essentielle lors de la phase d'apprentissage (voir section II.1.3).

Les protocoles d'entraînement nécessaires (détaillés dans la section II.3.2) sont bien souvent longs, pouvant mener à une perte de motivation et un découragement de la part des utilisateurs potentiels (Hamilton-Flectcher et al., 2016b). Ici encore, les environnements virtuels et augmentés offrent un potentiel intéressant en permettant des méthodes d'entraînement plus ludiques, toujours en impliquant des actions motrices volontaires accélérant l'apprentissage. Ils offrent également la possibilité d'adapter le protocole d'entraînement en fonction des progrès, en modulant progressivement la complexité de l'environnement, ce qui favorise la généralisation de l'apprentissage à des situations nouvelles (Maidenbaum et al., 2014a).

Une autre difficulté pouvant rendre ces systèmes difficiles d'accès est la nécessiter de constituer, lors de l'apprentissage, de nouvelles associations entre des sensations auditives et des éléments distants, en fonction du schéma d'encodage du DSS faisant correspondre des dimensions graphiques à des indices acoustiques. La non-intuitivité des schémas d'encodage utilisés dans les DSSs a été rapportée comme rendant difficile leur interprétation (Hamilton-Fletcher et al., 2016b) et constituant un frein à leur adoption, révélant l'importance de déterminer des indices acoustiques qui soient rapidement interprétables. Puisque l'apprentissage repose en partie sur les capacités

perceptives et cognitives préexistantes (voir section II.1.2), il est essentiel de développer un DSS vision-vers-audition en prenant en compte les capacités humaines cognitives et de perception. Les indices acoustiques utilisés dans le schéma d'encodage doivent également être déterminés au regard de la fonctionnalité visée du DSS (e.g., la localisation d'obstacles). D'une part, cela passe par la prise en compte des informations visuelles nécessaires à la tâche en question, et d'une autre part par l'adéquation des indices acoustiques associés à ces informations visuelles (Auvray, 2019 ; Loomis et al., 2018). Par exemple, dans le cadre d'un DSS pour l'aide à la locomotion et à la localisation d'obstacles, le champ visuel et la résolution graphique nécessaire à cette tâche doivent être pris en compte. Mais la détermination des indices acoustiques doit aussi, quant à elle, prendre en compte des aspects perceptifs telle que la résolution temporelle et spatiale de la modalité auditive (voir section II.2)

Enfin, l'évaluation des DSSs ne se fait pas systématiquement auprès de la population cible non-voyante, mais bien souvent auprès d'une population voyante à qui la vision est momentanément retirée avec un bandeau sur les yeux. Or, la population cible non-voyante peut se distinguer de façon importante de la population voyante en termes de capacités de perception auditive et spatiale (voir sections II.2.1.2.4, II.2.3.5 et II.2.4.3) et elle est de plus caractérisée par une forte hétérogénéité qui dépend notamment de la présence ou non d'une expérience visuelle antérieure (Arnold et al., 2017; Auvray, 2019). Bien entendu, évaluer les performances d'utilisation d'un DSS auprès de personnes voyantes ayant les yeux fermés permet d'apporter une preuve de concept du dispositif qui semble essentielle (Buchs et al., 2021). Néanmoins, cet aspect méthodologique soulève l'importance de développer des protocoles d'évaluation et d'entraînement qui puissent être adaptés à la population cible non-voyante.

1.5. Synthèse

- Les dispositifs de substitution sensorielle (DSS) sont des systèmes permettant de percevoir des informations initialement perçues à travers un canal sensoriel par le biais d'une autre modalité sensorielle. Ils visent à supplémenter les personnes non-voyantes en les aidant à percevoir et interagir avec leur environnement proche. Les DSSs vision-vers-audition convertissent les informations visuelles en des informations auditives en mettant en correspondance les caractéristiques graphiques d'une image avec des indices acoustiques.

- Les DSSs vision-vers-audition peuvent être utilisés dans des tâches de reconnaissance d'objet et de perception spatiale comme par exemple des tâches de localisation d'objets ou des tâches de navigation.

- Les DSSs peuvent être considérés comme des dispositifs d'extension de perception. D'après la théorie de l'intégration verticale, la perception avec un DSS repose sur des capacités préexistantes cognitives et perceptives.

- D'après la théorie de la hiérarchie inverse, la maîtrise d'un DSS nécessite un apprentissage perceptif durant lequel les liens entre les représentations de bas niveaux de traitement et de hauts niveaux de traitement sont ajustés pour permettre l'intégration de nouveaux indices acoustiques sous la forme de représentations de haut niveau ayant du sens.

- À terme du processus d'apprentissage, les sensations proximales auditives provenant du DSS sont automatiquement attribuées à des perceptions extérieures. Cette attribution distale est possible lorsque des actions motrices volontaires permettant d'acquérir les contingences sensorimotrices ont été employées.

- Parmi les freins à l'adoption des DSSs par la population cible non-voyante, on retrouve la surcharge cognitive et perceptive induite par les difficultés à interpréter suffisamment efficacement les nouveaux indices acoustiques pour effectuer des tâches, et la pénibilité des entraînements.

Dans le contexte du développement d'un DSS vision-vers-audition pour l'aide à la locomotion et à la localisation d'obstacles, il est nécessaire de prendre en compte les capacités préexistantes de perception auditive spatiale pour déterminer les indices acoustiques à moduler, afin de faciliter l'apprentissage et limiter la surcharge cognitive et perceptive. La section suivante présente les indices acoustiques spatiaux utilisés pour localiser une source sonore ainsi que les capacités humaines et limites de perception spatiale auditive.

---

2. Perception spatiale auditive

2.1. Perception auditive : d'un signal acoustique à un percept sonore

Les sons résultent de variations de la pression dans un milieu de propagation, par exemple l'air. Ces variations de pression proviennent principalement des vibrations résultantes d'une perturbation mécanique sur un corps matériel. Les vibrations mécaniques créent des variations locales de pression dans l'environnement : les ondes sonores. Lorsque le milieu de propagation est l'air, les ondes sonores se propagent à une vitesse d'environ 344 m/sec (avec une température de 20°, une pression atmosphérique de 100 kPa et une humidité de 50%). Les ondes sonores peuvent alors être perçues par le système auditif à travers deux organes récepteurs, les oreilles.

2.1.1. Anatomie du système auditif

Le système auditif humain est composé de l'oreille externe, moyenne, et interne. Chaque section anatomique de ce système perceptif va activement participer à la perception auditive. L'oreille externe, composée du pavillon et du canal auditif va jouer un rôle essentiel dans la perception auditive, notamment spatiale, en modulant la composition spectrale du son et en faisant converger le signal acoustique vers le tympan. Elle joue un rôle d'amplificateur et d'atténuateur de certaines fréquences à travers sa propre fonction de transfert. Au fond du canal auditif, se trouve le tympan qui sépare l'oreille externe de l'oreille moyenne et qui joue un rôle de conversion des ondes de pression en vibrations mécaniques qui se propagent dans l'oreille moyenne. L'oreille moyenne, composée des trois os que sont le marteau, l'enclume et l'étrier, joue un rôle de propagation des vibrations mécaniques du tympan jusqu'à l'oreille interne.

Dans l'oreille interne, se trouve la cochlée, organe récepteur en forme de spirale, sur laquelle s'étale la membrane basilaire sur laquelle repose l'organe de Corti contenant des cellules réceptrices que sont les cellules ciliées. Les vibrations mécaniques communiquées par l'oreille moyenne atteignent la partie basale de la cochlée et se propagent jusqu'à l'apex. La membrane basilaire est caractérisée par une variation de sa largeur, avec un élargissement et un épaississement au fur et à mesure de l'approche de l'apex. Cette particularité anatomique lui confère une propriété d'organisation tonotopique permettant un premier traitement de la dimension fréquentielle du signal acoustique. Agissant comme un filtre fréquentiel, elle répond préférentiellement à des fréquences hautes sur la partie basale (étroite et fine) et se spécialise dans des fréquences de plus en plus basses en allant vers la partie de l'apex (large et épaisse). La cochlée peut alors être subdivisée en régions qui répondent préférentiellement à certaines gammes de fréquence : les bandes critiques (Moore & Glasberg, 1983 ; Zwicker, 1961). L'énergie mécanique des vibrations

des cellules ciliées est ensuite convertie en signal nerveux, au niveau de l'organe de Corti, signal qui est ensuite transmis par le nerf auditif vers les structures cérébrales impliquées dans son traitement.

2.1.2. Multidimensionnalité d'un son

Un son est un percept multidimensionnel. Le signal acoustique peut être caractérisé par plusieurs éléments physiques influençant sa perception. Tout signal acoustique peut être caractérisé par une combinaison d'ondes sinusoïdales à des fréquences et amplitudes spécifiques, et caractérisées chacune par une phase. Chaque onde sinusoïdale s(t, f, A, φ) composant le signal acoustique est définie selon la fonction suivante :

$$s(t, f, A, \varphi) = A \sin(2\pi f t + \varphi),$$

avec t le temps, A l'amplitude, f fréquence et φ la phase.

Tout signal acoustique S(t) peut alors être décomposé sous le forme suivante :

$$S(t) = \sum_i s_i(t, f, A, \varphi) = \sum_i A_i \sin(2\pi f_i t + \varphi_i)$$

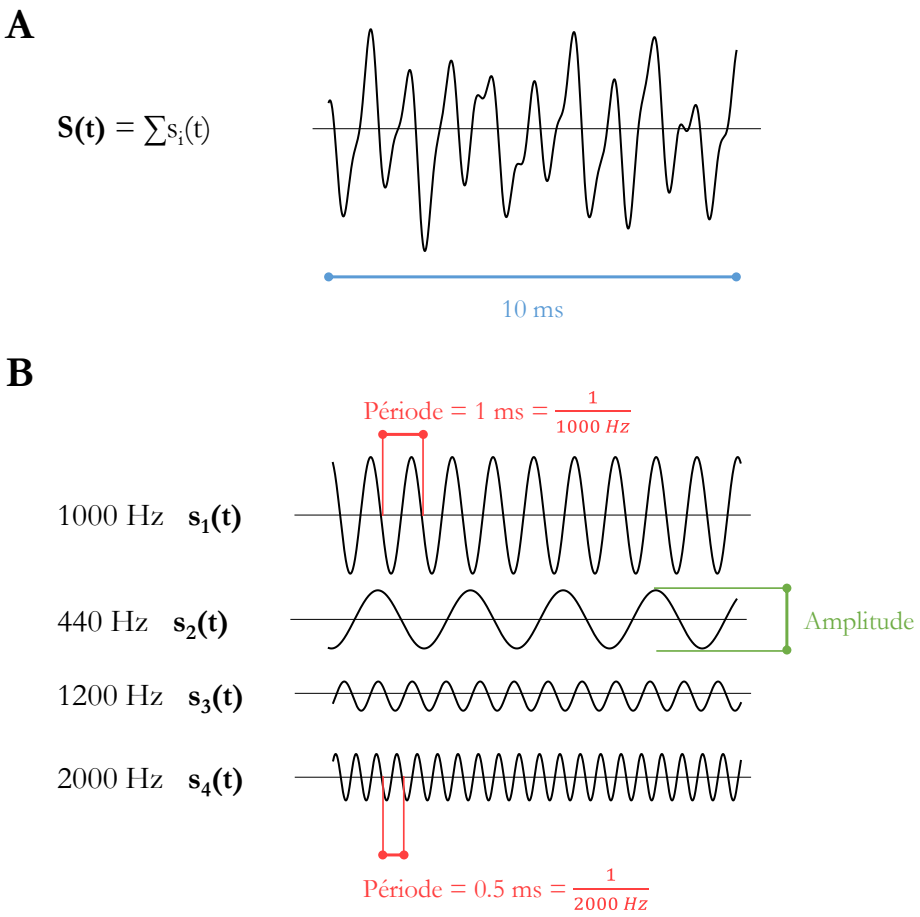
En conséquence, il est possible de générer un signal acoustique complexe en additionnant plusieurs signaux acoustiques purs : c'est la synthèse sonore additive. Un exemple de signal acoustique S(t) de 10 ms généré avec une méthode de synthèse additive est illustré dans la **Figure II-3**. Le signal acoustique S(t) est généré à partir d'une combinaison de 4 ondes sinusoïdales pures (1000 Hz, 440 Hz, 1200 Hz et 2000 Hz) dont l'amplitude relative varie (2, 1, 0.5 et 0.8) et dont les phases sont aléatoires.

2.1.2.1. Fréquence et tonie

2.1.2.1.1. Sons purs

Le signal acoustique peut être décomposé en une combinaison d'ondes sinusoïdales, chacune oscillant à une certaine fréquence et caractérisée par une phase. La fréquence du signal acoustique, exprimée en Hertz, correspond au nombre de cycles réalisés pendant une seconde. Par exemple, le signal acoustique pur $s_1(t)$ oscillant à 1000 Hz (**Figure II-3** B, haut) complète 10 cycles en 10 ms (soit 1000 cycles en 1 seconde). Sa période, soit la durée d'un cycle, est donc de 1 ms. Pour un son pur, une seule fréquence est représentée dans le signal acoustique. Chez l'adulte typique, le spectre audible par le système auditif s'étend de 20 Hz à environ 16000 Hz, bien que les capacités de perception puissent s'étendre à de plus hautes fréquences jusqu'à 20000 Hz à un plus jeune âge.

Figure II-3. Exemple d'un signal acoustique complexe S(t) d'une durée de 10 ms **(A)** généré par synthèse additive à partir d'une combinaison de 4 ondes sinusoïdales $s_i(t)$ **(B)** oscillant à la fréquence indiquée à gauche du tracé.

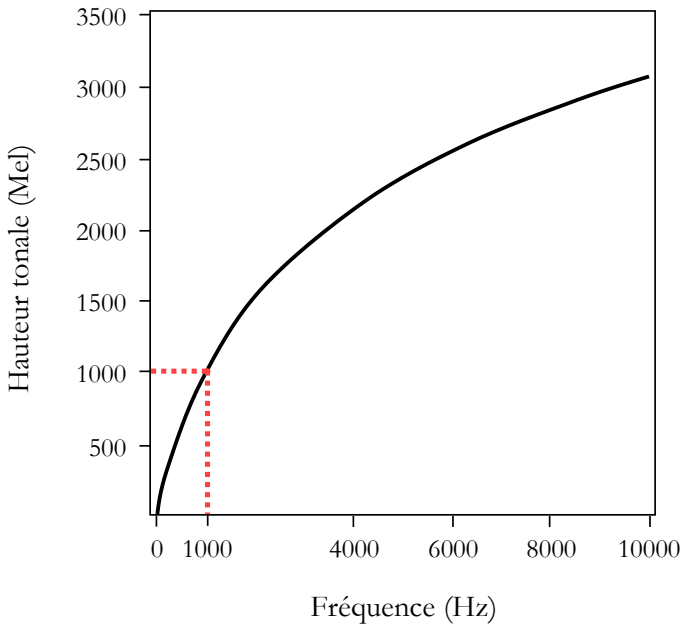


La tonie (ou hauteur tonale) d'un son correspond à la caractéristique perceptive qui permet de qualifier un son sur une échelle de grave à aigu. Dans le cas d'une tonalité pure, c'est-à-dire un son dont une seule fréquence est représentée, la fréquence du signal est à l'origine du percept de tonie (Plack & Oxenham, 2006). Plus la fréquence du signal acoustique est élevée, plus la tonie l'est, mais la relation entre fréquence et tonie n'est pas linéaire, mais plutôt logarithmique. Plusieurs échelles représentent la relation entre fréquence du signal acoustique et la hauteur tonale : l'échelle des Mel (Stevens et al., 1937), l'échelle des Barks (Zwicker, 1961), l'échelle des ERB (Moore & Glasberg, 1983). Parmi elles, l'échelle des Mel (**Figure II-4**) approxime la hauteur tonale (en Mel) perçue pour une onde sinusoïdale pure de fréquence f (en Hz) avec la formule suivante :

$$Mel = 2595 \times log_{10}\left(1 + \frac{f}{700}\right)$$

Avec l'échelle de Mel, une fréquence de 1000 Hz équivaut à 1000 Mel, mais une fréquence de 250 Hz équivaut à 344 Mel, et une fréquence de 1500 Hz équivaut à 1291 Mel.

Figure II-4. Echelle des Mel. Fréquence en Mel en fonction de la fréquence physique du signal acoustique en Hertz.

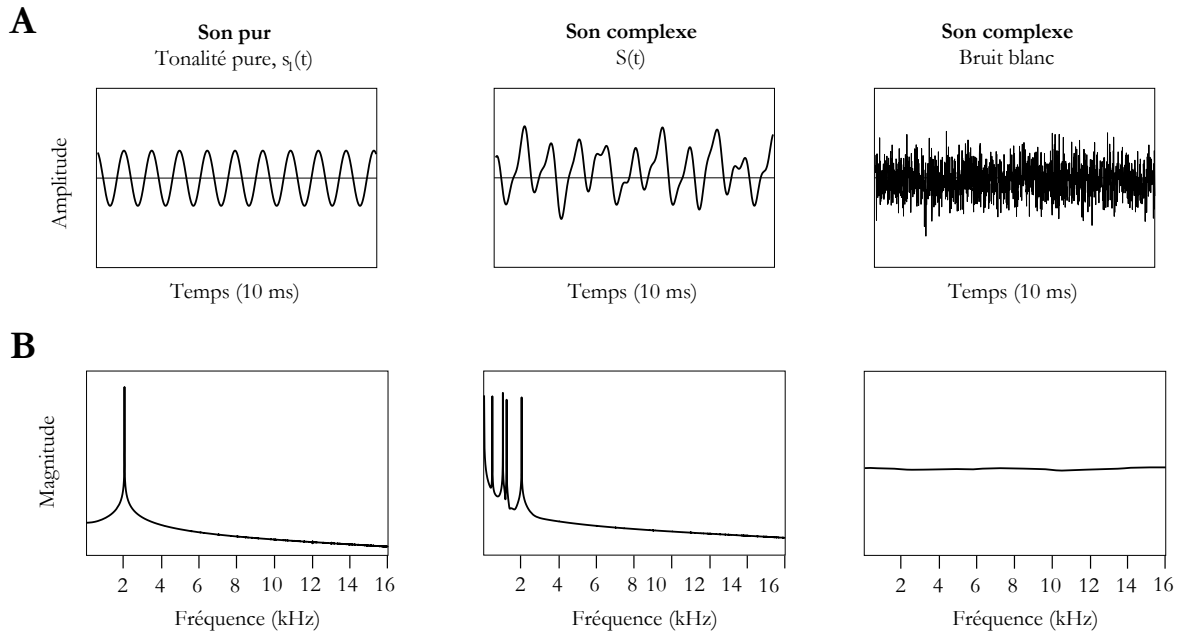


Si le spectre de fréquence perceptible s'étend entre 20 et 16 000 Hz, les capacités de perception du système auditif sont hétérogènes sur cet intervalle. Elles sont meilleures sur l'intervalle de fréquence allant de 1500 Hz à 6000 Hz et optimales entre 3000 et 4000 Hz (Scharine et al., 2009). Pour des tonalités pures, les capacités de discrimination de deux fréquences entre 200 et 2000 Hz sont de l'ordre de 1 à 3 Hz (Wier et al., 1977). Au-delà, il est nécessaire que la différence en fréquence entre deux tonalités soit supérieure pour qu'une différence de tonie soit perçue.

2.1.2.1.2. Sons complexes

Au quotidien, le système auditif traite des sons bien plus complexes que des tonalités pures, c'est-à-dire des sons comportant plusieurs composantes fréquentielles. Alors que la complexité spectrale correspond à la quantité de composantes fréquentielles contenues dans le signal acoustique, sa largeur de bande définit l'étendue des fréquences le composant. Plus le spectre d'un son est complexe, plus il contient de fréquences différentes. Plus la largeur de bande du spectre d'un son est grande, plus il est composé de fréquences éloignées. La **Figure II-5** illustre la complexité spectrale de trois signaux acoustiques de différentes complexités et largeurs de bande. La tonalité pure $s_1(t)$ oscillant à 1000 Hz est la moins complexe et a une largeur de bande très étroite. Un premier son complexe S(t) composé de quatre fréquences (440 Hz, 1000 Hz, 1200 Hz et 2000 Hz) est caractérisé par une largeur de bande de 1560 Hz (440–2000 Hz), et une complexité spectrale de l'ordre de 4 fréquences. À l'inverse, le bruit blanc est un signal acoustique dont les fréquences perceptibles (20 à 20 000 Hz) sont équitablement représentées dans le spectre. Il s'agit donc d'un signal acoustique à large bande et haute complexité spectrale.

Figure II-5. Onde **(A)** et spectre de fréquences associé **(B)** de trois signaux acoustiques de différentes complexités.



Les propriétés physiques des matériaux font que très souvent, un objet produit des vibrations acoustiques à la suite de la rentrée en résonance de sa structure interne. Dans ce cas, le son produit est quasi-périodique et le système auditif perçoit une unique hauteur tonale même si le signal acoustique est composé de plusieurs fréquences : une fréquence fondamentale et des fréquences harmoniques. La fréquence fondamentale est la plus petite fréquence alors que les harmoniques correspondent aux multiples de la fréquence fondamentale. Pour illustrer ces propos, prenons une fréquence fondamentale de 1000 Hz. Son premier harmonique est à 2000 Hz (2×1000 Hz), son troisième est à 3000 Hz (3×1000 Hz), et ainsi de suite. Ainsi, le $n^{ème}$ harmonique d'une fréquence fondamentale f_0 correspond à $f_0 \times n$ Hz. Pour cette même fréquence fondamentale donnée f_0 =1000 Hz, elle augmentera d'une octave à chaque fois que sa fréquence doublera. Par exemple, il y a une octave entre 1000 Hz et 2000 Hz (2×1000 Hz), et également une octave entre 2000 Hz et 4000 Hz (2×2000 Hz). Ainsi, pour une fréquence donnée f_0, on pourra calculer sa $i^{ème}$ octave avec la formule suivante $f_0 \times 2^i$. D'un point de vue perceptif, l'octave est à l'origine de la consonance, c'est-à-dire la notion d'harmonie, ou de plaisance, à l'écoute d'un son (D. M. Howard & Angus, 2009).

Au-delà de la perception de la tonie, la composition fréquentielle d'un signal acoustique joue un rôle dans la sensation de désagréabilité qui peut être évoquée lors de l'écoute d'un son. Comme le rapporte Kumar et al. (2008), les sons dont l'intervalle de fréquence est compris entre 2500 et 5500 Hz ont tendance à être jugés comme désagréables.

2.1.2.2. Intensité et sonie

D'un point de vue acoustique, l'intensité correspond à la magnitude de la variation de pression dans le milieu de propagation. L'intensité moyenne d'un signal acoustique est proportionnelle à la moyenne du carré de la pression acoustique enregistrée. Chez un jeune adulte typique sans trouble auditif, une variation de la pression acoustique d'au moins 20 µPa est souvent nécessaire pour qu'un son soit perçu : c'est le seuil d'audition. Par praticité, on utilise l'unité dB SPL (pour *Sound Pressure Level*), qui correspond à la formule suivante :

$$\text{dB SPL} = 20 \times log_{10}\left(\frac{p_{mesurée}}{p_{référence}}\right)$$

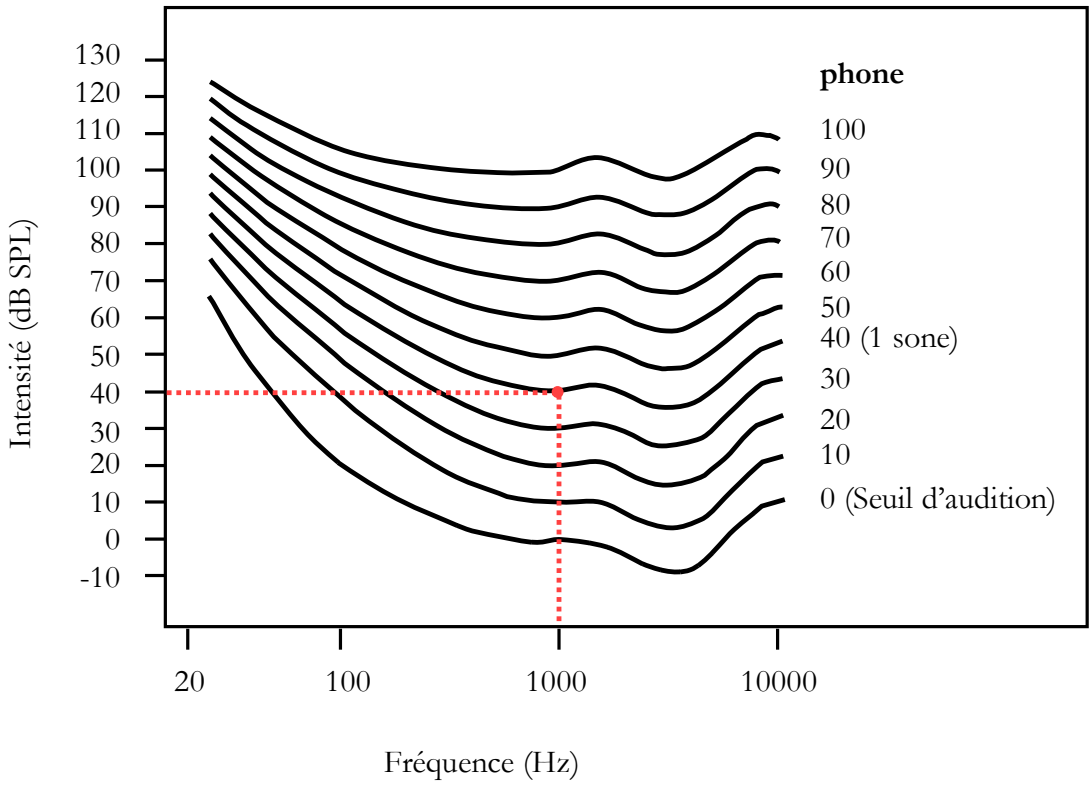
avec $p_{référence}$ = 20 µPa la pression sonore associée au seuil de perception auditive, et $p_{mesurée}$ la pression sonore mesurée. Ainsi, 0 dB SPL est associé au seuil de perception auditive.

La sonie correspond à la perception de l'intensité du signal acoustique, c'est-à-dire à la sensation de force sonore induite par le son, allant de faible à forte. La sonie se mesure avec l'unité de sone, sachant qu'une valeur de 1 sone correspond à la sensation de force sonore induite par une tonalité pure de 1000 Hz diffusée à 40 dB SPL. Ainsi, une valeur de 2 sones correspondra à une sensation de sonie deux fois plus forte que celle de 1 sone. La sonie dépend des propriétés physiques du signal acoustique, dont sa composition fréquentielle. Deux tonalités pures de différentes fréquences diffusées avec un même niveau de pression acoustique ne seront pas perçues avec la même sonie. L'unité de mesure phone exprime la sonie perçue indépendamment de la fréquence. Elle est définie telle que deux tonalités de fréquences distinctes ayant le même niveau en phone sont perçues avec la même sonie. Par convention, 1 sone correspond à une mesure de 40 phones, c'est-à-dire à la sensation de force sonore induite par une tonalité pure de 1000 Hz diffusée à 40 dB SPL, et 0 phone correspond à 0 dB SPL (**Figure II-6**). Pour des tonalités pures, l'équivalence de sonie entre les tonalités est représentée par un niveau d'isosonie, associé à une unité de phone. Pour un niveau d'isosonie donné (e.g., 40 phones), on peut alors déterminer l'intensité en dB SPL à utiliser pour chaque fréquence pour que la sensation de force (la sonie) soit équivalente pour chacune des fréquences. La norme ISO 226 :2003 propose une approximation des courbes isosoniques pour différentes phonies (**Figure II-6**).

Les capacités de discrimination de l'intensité du système auditif dépendent de la composition fréquentielle des sons. Elles sont de l'ordre de 0.4 dB pour des sons à large bande (G. A. Miller, 1947), et entre 1 et 2 dB pour des tonalités pures allant de 200 à 8000 Hz (Jesteadt et al., 1977). De façon générale, les capacités de discrimination d'intensité diminuent à mesure que l'intensité augmente (Jesteadt et al., 1977). En d'autres termes, plus l'intensité du son est forte, plus une importante augmentation de l'intensité est nécessaire pour qu'elle soit détectée.

Comme vu précédemment, l'organisation tonotopique de la cochlée lui confère une propriété de filtre fréquentiel (voir section II.2.1.1). Lorsqu'une tonalité pure atteint la cochlée, une région spécifique de la membrane basilaire répond à cette fréquence. Cette région de la membrane répond préférentiellement à un intervalle de fréquence, c'est-à-dire à une bande critique telle que définie par Zwicker (1961). D'un point de vue perceptif, cela peut se traduire par un effet de masquage lorsque deux tonalités oscillant à deux fréquences distinctes, mais proches (i.e., deux tonalités stimulant la même région de la membrane basilaire) sont diffusées simultanément. Dès lors que les fréquences des tonalités oscillent à deux fréquences suffisamment éloignées, elles stimulent deux régions suffisamment distinctes de la membrane basilaire, et l'effet de masquage prend fin et la sonie augmente (Glasberg & Moore, 1990).

Figure II-6. Courbes isosoniques définies par la norme ISO 226 :2003. Chaque courbe indique l'intensité nécessaire en dB SPL pour obtenir une sonie équivalente entre les fréquences données (valeurs à droite, en phone). Une valeur de 40 phone correspond à une intensité de 40 dB SPL pour une tonalité pure oscillant à 1000 Hz (rouge).



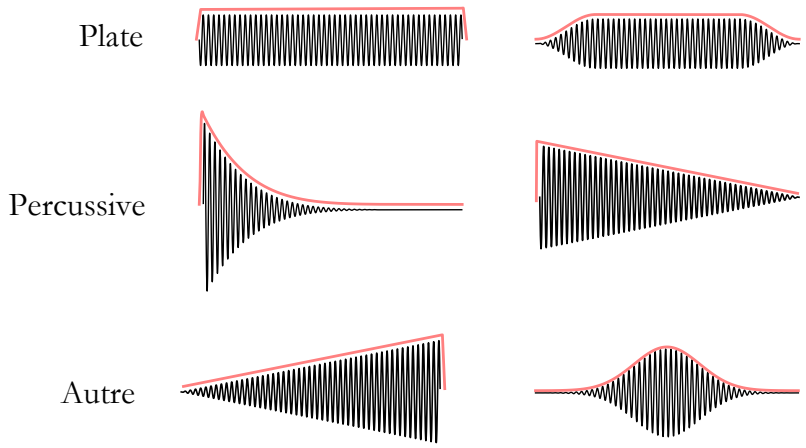
2.1.2.3. Enveloppe et timbre

Un signal acoustique peut être caractérisé par son enveloppe qui correspond à la distribution de l'énergie au cours du temps à très basses fréquences. Dans la synthèse acoustique, on peut décomposer l'enveloppe en trois phases : le temps d'attaque (*onset* en anglais), la période de maintien (*sustain* en anglais), et le relâchement (*offset* en anglais) (Begault, 1995). Le temps d'attaque correspond à la première période du son, c'est-à-dire la phase durant laquelle l'énergie

dans le signal acoustique augmente. Elle peut être suivie d'une période de maintien durant laquelle l'énergie dans le signal reste stable. Enfin, le temps de relâchement est associé à la diminution de l'énergie dans le signal jusqu'à ce qu'il devienne nul à nouveau.

À partir de cette caractérisation en 3 phases, on peut différencier différents types d'enveloppes, telles que les enveloppes plates et les enveloppes percussives (Schutz & Gillard, 2020). Les enveloppes plates sont caractérisées par la présence d'une période de maintien (**Figure II-7**, première ligne), alors que les temps d'attaque et de relâchement peuvent être abrupts ou progressifs. À l'inverse, les enveloppes percussives (**Figure II-7**, deuxième ligne) sont caractérisées par un temps d'attaque très abrupte, suivi d'un temps de décroissance rapide. Elles ne contiennent donc pas de période de maintien. D'autres catégories d'amplitudes d'enveloppe sans période de maintien peuvent être décrites, mais avec un temps d'attaque plus long. C'est le cas des enveloppes Gaussiennes (**Figure II-7**, troisième ligne, à gauche) ou des enveloppes avec une attaque linéaire plus longue suivie d'un temps court de relâchement (*Rising* en anglais) (**Figure II-7**, troisième ligne, à droite).

Figure II-7. Exemples d'amplitudes d'enveloppe appliquées sur une tonalité pure de fréquence 500 Hz et d'une durée de 100 ms classées par catégories (Plate, Percussive et Autre). La forme de l'enveloppe supérieure est tracée en rouge.



L'amplitude de l'enveloppe influence plusieurs attributs perceptifs du son. Elle est intimement impliquée dans la perception du timbre, une dimension du son qui contribue (avec le spectre harmonique) à différencier une même tonalité jouée avec deux instruments de musique tels qu'un piano et un violon. Dans le contexte de synthèse acoustique de stimuli simples tels que des tonalités pures, l'amplitude de l'enveloppe est par exemple impliquée dans la perception d'un son de nature percussive (Schutz & Gillard, 2020), qui est aussi évoqué au quotidien lors d'impacts entre des objets tels qu'une bouteille en verre tombant au sol. La perception d'un clic à la place d'une tonalité pour des sons de très courtes durées (inférieures à 25 ms) peut être influencée par l'enveloppe, du fait des modifications induites sur le spectre de fréquence. Par exemple, Mohlin

(2011) met en évidence que des tonalités dont la période de décroissance de l'enveloppe est modulée exponentiellement nécessitent d'être d'une durée plus longue qu'avec des enveloppes gaussiennes pour qu'une tonalité soit perçue au lieu d'un clic.

Au-delà du timbre, l'amplitude de l'enveloppe peut influencer des attributs tels que la durée d'un son (Vallet et al., 2014), sa sonie (Neuhoff, 1998 ; Ries et al., 2008) et sa tonie (Moore, 2008 ; Rossing & Houtsma, 1986). Par exemple, les tonalités percussives (**Figure II-7**, deuxième ligne) ont tendance à être perçues comme étant de plus courtes durées que les tonalités avec une amplitude d'enveloppe plate (**Figure II-7**, première ligne). Avec des sons complexes à larges bandes, Ries et al. (2008) ont montré que lorsque l'amplitude de l'enveloppe était croissante, les sons avaient tendance à être perçus de plus longue durée et plus forte sonie que les sons dont l'enveloppe était décroissante. Avec des tonalités pures et complexes, Neuhoff (1998) a également montré que la sonie était supérieure lorsque l'amplitude de l'enveloppe était croissante plutôt que décroissante. Avec des tonalités pures, Rossing & Houtsma (1986) démontrent qu'une modulation exponentielle de l'enveloppe peut induire un décalage dans la hauteur tonale, la tonie étant majoritairement perçue comme plus aigüe.

2.1.2.4. Perception auditive chez les personnes non-voyantes

La privation visuelle chez les personnes non-voyantes entraîne des mécanismes de compensation pouvant moduler les capacités de perception auditive, telles que la capacité à percevoir des changements de fréquence ou de timbre, ou à percevoir un son diffusé dans un fond sonore environnant.

Comparativement à des personnes voyantes, de meilleures capacités de discrimination de fréquences ont été mesurées chez des personnes non-voyantes précoces lorsque des sons purs étaient utilisés (Arnaud et al., 2018 ; Gougoux et al., 2004 ; Rokem & Ahissar, 2009 ; Wan et al., 2010). Par exemple, Wan et al. (2010) rapportent que pour des tonalités pures entre 500 et 1500 Hz les participants non-voyants discriminaient mieux que les participants voyants deux tonalités séparées de 0.25 % à 2 % Hz, ce que rapporte Gougoux et al. (2004) dans une autre étude. En utilisant des tonalités pures entre 1000 et 2000 Hz, Rokem et Ahissar (2009) ont mesuré des seuils de discrimination de fréquence plus bas chez les personnes non-voyantes, ce qu'observe aussi Arnaud et al. (2018) avec des fréquences plus basses de l'ordre de 100 Hz. De meilleures capacités de discrimination de fréquences ont également été mesurées chez des personnes non-voyantes lorsque des sons complexes étaient utilisés tels que des voyelles ou des notes jouées avec un instrument de musique (Arnaud et al., 2018), ou des tonalités complexes avec harmoniques (Wan et al., 2010).

La population des personnes non-voyantes est caractérisée par une forte hétérogénéité intrinsèquement due à la présence ou non d'une expérience visuelle passée et de sa durée. En comparant les capacités de discrimination de fréquences, Gougoux et al. (2004) et Wan et al (2010) ont mesuré de meilleures capacités de discrimination chez les non-voyants précoces que chez les non-voyants tardifs.

Ces meilleures capacités de perception auditives constatées chez les personnes non-voyantes ne se restreignent pas à la perception des fréquences, mais également à la discrimination du timbre, comme suggéré par Wan et al. (2010). Dans cette étude, le timbre d'une tonalité complexe était manipulé en modulant la distance entre la fréquence fondamentale et les harmoniques, et les participants, voyants et non-voyants, devaient indiquer si le timbre était différent. Ils rapportent que les participants non-voyants précoces avaient de meilleures capacités de discrimination du timbre comparativement voyants mais également aux non-voyants tardifs. En évaluant le seuil de perception d'un pseudo-mot diffusé parmi un fond sonore, Rokem et Ahissar (2009) mesurent quant à eux de meilleures capacités de perception d'un signal parmi le bruit chez les non-voyants que chez les voyants.

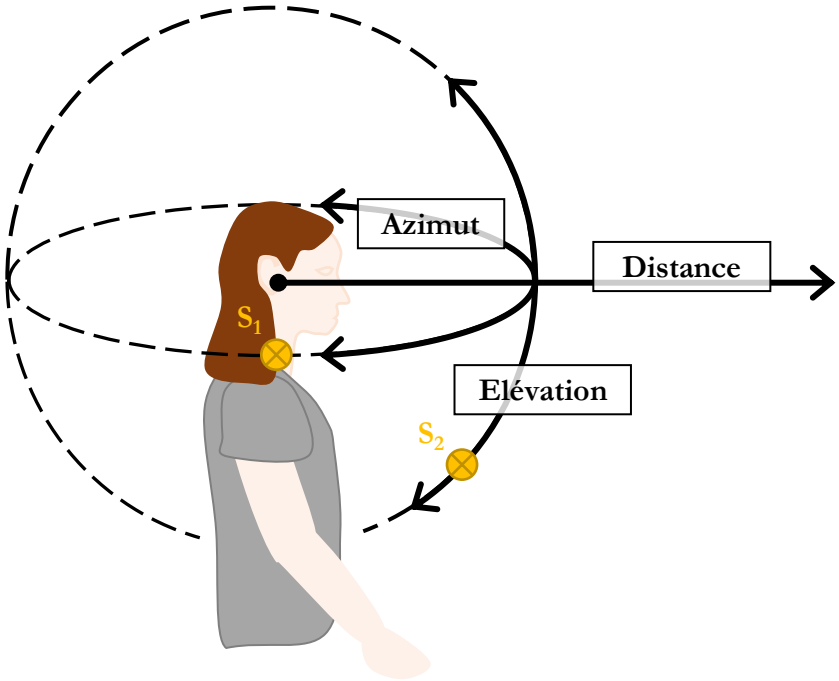
Finalement, ces études suggèrent la présence de mécanismes de compensation pouvant entraîner de meilleures capacités de perception auditive chez les personnes non-voyantes. Certains des mécanismes de compensation arriveraient précocement au cours du développement, comme le suggèrent les meilleures capacités observées chez les personnes non-voyantes précoces que chez les personnes non-voyantes tardives.

2.2. Indices acoustiques pour localiser une source sonore

La localisation d'une source sonore est perçue dans un espace en 3 dimensions dont le cadre de référence en coordonnées sphériques est centré sur la tête de l'auditeur, à hauteur des organes récepteurs (i.e., les oreilles) (**Figure II-8**). Les coordonnées sphériques se déclinent en coordonnées angulaires pour la direction (azimut et élévation), et coordonnées linéaires pour la distance. L'azimut correspond à la latéralité de la source sonore, allant de -180° à +180°, un azimut de 0° étant associé à une position le long de l'axe médian, c'est-à-dire centrée par rapport aux deux oreilles. Une valeur négative de l'azimut est associée à une source sonore positionnée à gauche de l'auditeur, alors qu'une valeur positive est associée à une source sonore localisée sur sa droite (e.g., S_1 est placée à un azimut de +90° dans la **Figure II-8**). L'élévation indique la hauteur de la source sonore et varie également entre -180° et +180°, avec 0° correspondant à une position à la hauteur des oreilles. Ainsi, une valeur négative en l'élévation correspond à une source sonore localisée plus bas que les oreilles (e.g., S_2 est placée à une élévation de -45° dans la **Figure II-8**), alors qu'une valeur positive indique une source sonore située plus haut que les oreilles.

La perception spatiale auditive repose sur plusieurs mécanismes. Durant sa propagation, le signal acoustique subit des effets de filtrage par l'environnement de propagation et le corps de l'auditeur. Pour percevoir la position d'une source sonore dans l'espace en 3 dimensions (azimut, élévation, et distance), le système auditif utilise des indices acoustiques qui peuvent être distingués en deux catégories : les indices binauraux et monauraux.

Figure II-8. Schématisation du système de coordonnées sphériques utilisé pour la localisation auditive. L'azimut et l'élévation sont des métriques angulaires alors que la distance est linéaire. L'azimut correspond à la position latérale de la source sonore, alors que l'élévation correspond à sa position verticale. Deux exemples de sources sonores (S_1 et S_2) localisées à une même distance de l'auditeur (0.5 m) sont présentés avec leurs coordonnées [azimut, élévation, distance], S_1 ayant pour position [+90°, 0°, 0.5 m], et S_2 la position [0°, -45°, 0.5 m].



Les indices acoustiques binauraux proviennent de la réception par les deux oreilles de signaux acoustiques différents même s'ils sont tous les deux issus de la même source sonore. Puisque les signaux acoustiques atteignent deux organes récepteurs distincts, le système auditif interprète les différences entre les signaux reçus au niveau de l'oreille droite et de l'oreille gauche. Les indices binauraux sont principalement utilisés pour percevoir la position en azimut d'une source sonore et reposent sur des comparaisons intéraurales temporelles et en intensité (détaillé dans la section II.2.2.1).

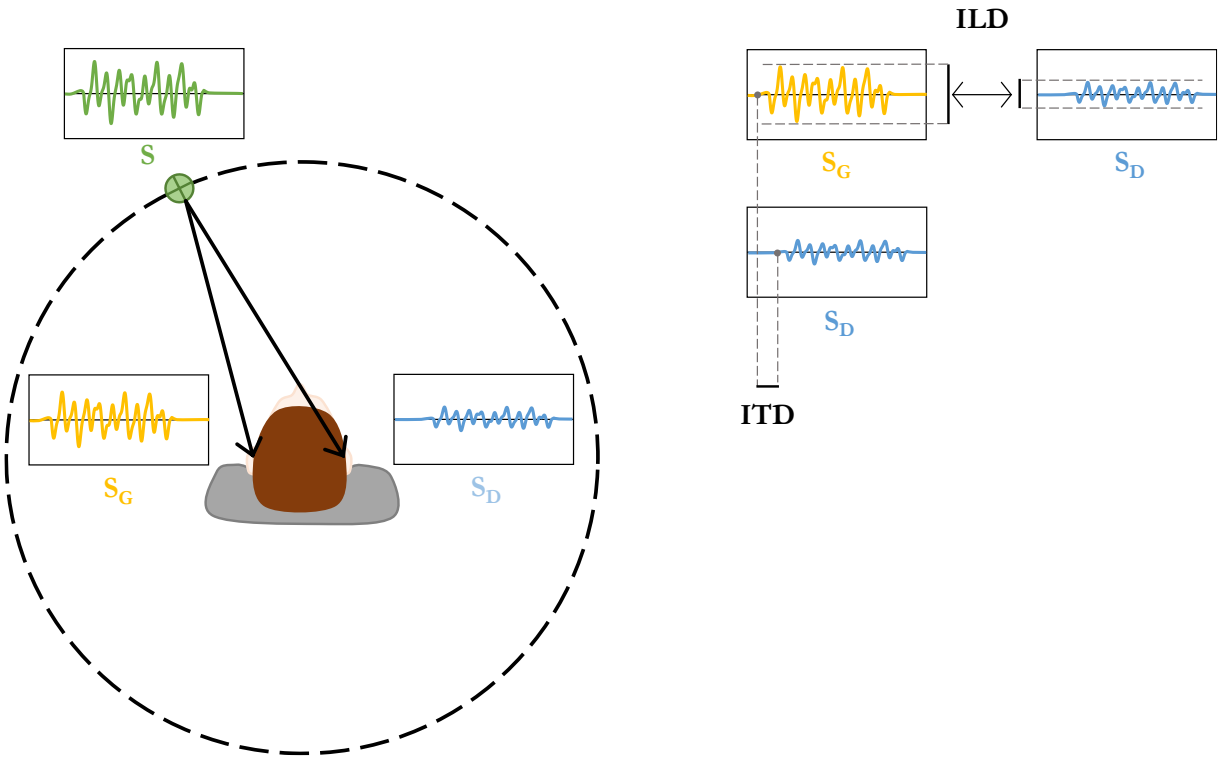
Les indices monauraux sont issus de la modification du signal acoustique lors de sa propagation entre la source sonore et les organes récepteurs. Ces altérations dépendent des caractéristiques de l'environnement de propagation et du corps de l'auditeur. Ces modifications, majoritairement spectrales, servent d'indices déterminants pour que le système auditif perçoive la

position en élévation (détaillé dans la section II.2.2.2) d'une source sonore et sa distance (détaillé dans la section II.2.2.3).

2.2.1. Indices pour l'azimut

Pour localiser la position en azimut d'une source sonore, le système auditif utilise des indices binauraux. Ces indices binauraux reposent sur une comparaison interaurale opérée entre les deux signaux acoustiques reçus. Lorsqu'un son est diffusé dans un hémichamp de l'auditeur (e.g., gauche sur la **Figure II-9**), les ondes sonores provenant de cette même source sonore atteignent à des moments différents l'oreille gauche et l'oreille droite (différence temporelle, *Interaural Time Difference*, ITD), et sont altérées différemment en amplitude (différence d'intensité, *Interaural Level Difference*, ILD). Le décalage intéraural temporel peut également induire un décalage de phase entre les signaux (différence de phase, *Interaural Phase Difference*, IPD). Le système auditif utilise ces trois indices binauraux pour estimer l'azimut de la source sonore. Comme stipulé dans la *Duplex theory* présentée par Rayleigh (1907), l'utilisation de ces trois indices acoustiques dépend du contenu fréquentiel de la source sonore : les indices temporels (ITD et IPD) prédominent pour les basses fréquences, alors que l'indice d'intensité (ILD) est prédominant pour les hautes fréquences.

Figure II-9. Localiser l'azimut d'une source sonore. La diffusion d'un signal acoustique S (vert) dans l'hémichamp gauche de l'auditeur se traduit par un décalage temporel (ITD) et d'intensité (ILD) entre le signal acoustique atteignant l'oreille ipsilatérale (S_G, jaune) et celui atteignant l'oreille controlatérale (S_D, bleu).



Prenons l'exemple d'un son diffusé à partir d'une source sonore localisée dans l'hémichamp gauche d'un auditeur (**Figure II-9** A). Sachant que les ondes sonores se propagent à une vitesse d'environ 344 m/s dans l'environnement de propagation, et que les organes récepteurs sont séparés spatialement, la variation de la pression acoustique est reçue plus précocement au niveau de l'appareil auditif ipsilatéral à la source que controlatéral (**Figure II-9** B). L'ITD traduit ce décalage temporal dû à la distance supplémentaire que l'onde sonore doit parcourir pour atteindre l'oreille controlatérale. En assumant que la tête de l'auditeur est sphérique, l'ITD peut être approximée par une relation entre la vitesse du son, l'angle d'arrivée (azimut) de l'onde sonore et la distance entre les deux oreilles de l'auditeur telle que :

$$ITD = \frac{r(\theta+sin\theta)}{v},$$

avec l'ITD exprimé (en seconde), r la moitié de la distance entre les deux oreilles (en mètre), θ l'azimut (en radians), et v la vitesse du son (en m/s) (D. M. Howard & Angus, 2009). En assumant que la moitié de la distance entre deux oreilles est de 9 cm (avec un diamètre crânien moyen de 18 cm), et que la vitesse du son est d'environ 344 m/s, on peut approximer l'ITD en fonction d'un azimut θ (en radian) par la formule suivante :

$$ITD(\theta) = \frac{0.09(\theta+sin\theta)}{344}.$$

Avec cette approximation, on obtient ainsi une ITD théorique maximum d'environ 673 µs pour un azimut de $\frac{\pi}{2}$ radians soit 90°. L'ITD varie en réalité en fonction de la fréquence du signal acoustique, avec des valeurs maximales d'environ 800 µs pour les basses fréquences et environ 600 µs pour les hautes fréquences (Benichoux et al., 2016). Le système auditif utilise l'ITD principalement pour localiser la position en azimut de sources sonores dont le contenu spectral comprend des fréquences inférieures à 1800 Hz (Benichoux et al., 2016 ; Blauert, 1983 ; Mills, 1958 ; Risoud et al., 2018), bien que cet indice acoustique puisse être utilisé pour des fréquences supérieures (Macpherson & Middlebrooks, 2002). La variation de l'ITD en fonction de l'azimut n'étant pas parfaitement symétrique par rapport à la tête, il s'agit d'un indice acoustique prédominant pour réussir à déterminer si une source sonore se trouve devant ou derrière l'auditeur (Benichoux et al., 2016).

Étant donné la nature souvent périodique des signaux sonores, le décalage intéraural temporel (ITD) est souvent associé à un décalage intéraural de phase (IPD). La différence de phase dépend de la période du signal acoustique (**Figure II-3** B), donc de sa fréquence, et peut être approximé par la formule suivante pour des tonalités pures :

$$IPD(\theta, f) = 2\pi \times f \times r(\theta + sin\theta),$$

avec f la fréquence en Hz, r la moitié de la distance entre les deux oreilles (en mètres), et θ l'azimut (en radians) (D. M. Howard & Angus, 2009).

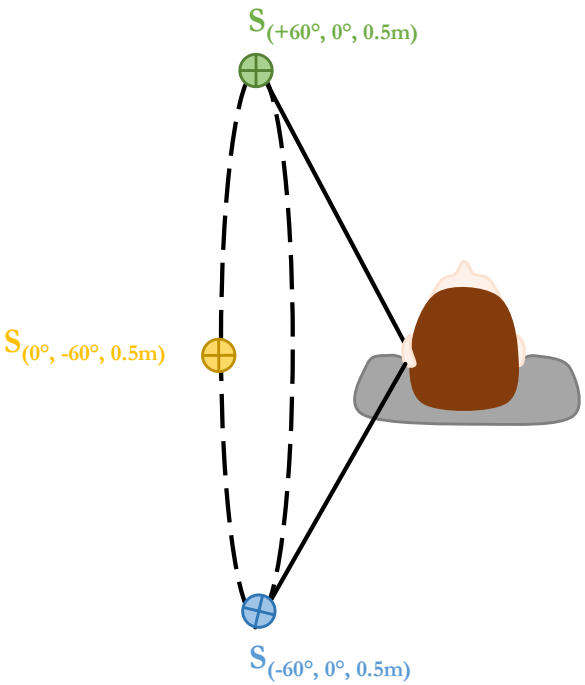
Les deux signaux acoustiques ne se différencient pas seulement par leur décalage temporel. En effet, lorsqu'un son est émis dans l'hémichamp gauche d'un auditeur, l'onde sonore entre en interaction avec la tête de l'auditeur avant d'atteindre le conduit auditif de l'oreille droite (**Figure II-9**, gauche), ayant pour effet d'atténuer l'amplitude du signal acoustique. Les signaux acoustiques reçus au niveau de l'oreille ipsilatérale et au niveau de l'oreille controlatérale à la source sonore auront donc tendance à avoir une différence d'intensité. L'ILD correspond à cet écart en intensité entre les signaux acoustiques atteignant l'organe récepteur gauche et droit, le signal acoustique arrivant à l'oreille controlatérale à la source sonore ayant tendance à s'atténuer davantage au cours de sa propagation (**Figure II-9**, droite). Il est plus difficile d'approximer l'ILD car elle est influencée par des phénomènes de diffraction et de réflexion sur des parties du corps (Akeroyd, 2014). L'ILD est un indice acoustique prédominant pour localiser l'azimut de sources sonores dont le contenu spectral comprend des fréquences supérieures à 2000 Hz (D. M. Howard & Angus, 2009). En dessous de 2000 Hz, la longueur d'onde (~ 17cm) est supérieure à la taille de la tête de l'auditeur, limitant l'effet d'atténuation du signal acoustique. Dans les faits, l'utilisation de l'ILD est optimale pour les fréquences supérieures à 3000 Hz (Risoud et al., 2018).

2.2.2. Indices pour l'élévation

Les indices binauraux qui sont utilisés pour localiser l'azimut d'une source sonore ne sont pas suffisants pour localiser la position en élévation. En effet, une multitude de positions distinctes dans l'espace peut être associée à des différences intéraurales d'intensité et temporelles similaires (ILD et ITD). Ces ensembles de positions sont appelés cônes de confusion **(Figure II-10)** et rendent insuffisants les indices binauraux pour discriminer la position de la source sonore le long d'un cône de confusion. Ainsi, pour localiser la position en élévation d'une source sonore, le système auditif se repose principalement sur des indices monauraux de nature spectrale. Au cours de sa propagation, le signal acoustique subit des modifications dans l'environnement de propagation, notamment induites par le corps de l'auditeur. Les ondes sonores entre en interaction avec des parties du corps de l'auditeur, notamment le pavillon de l'oreille, mais aussi la tête et le tronc. Ces interactions physiques entraînent des modifications de nature spectrale et comprennent des phénomènes de réflexion, diffraction, et absorption (Blauert, 1983). Pour un auditeur donné, des caractéristiques spectrales permettent au système auditif de déterminer la position en élévation d'une source sonore en associant ces caractéristiques spectrales à des positions dans l'espace. Les modifications spectrales se caractérisent par la combinaison de trous spectraux (*notches* en anglais)

et d'amplification de l'énergie dans certaines gammes de fréquences en fonction de l'élévation de la source sonore, qui reflètent principalement le filtrage fréquentiel du pavillon.

Figure II-10. Cône de confusion. Les différences intéraurales d'intensité et temporelles associées aux trois sources sonores $S_{(+60°, 0°, 0.5m)}$ (vert), $S_{(0°, -60°, 0.5m)}$ (jaune), et $S_{(-60°, 0°, 0.5m)}$ (bleu), positionnées sur le cône sont similaires. Les indices binauraux ne suffisent pas pour estimer la position des sources sonores sur le cône de confusion.



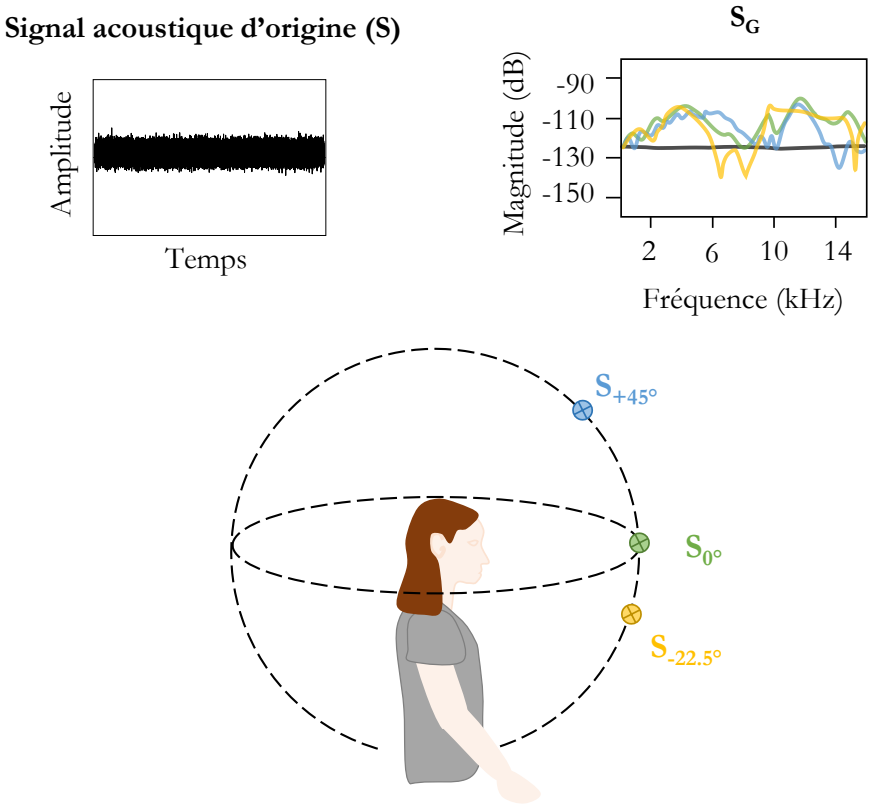
Puisque les modifications spectrales sont induites par des irrégularités morphologiques des pavillons, elles varient d'un individu à un autre. Néanmoins, elles ont le point commun d'être prédominantes au-delà de 4000 Hz (Algazi et al., 2001a ; Asano et al., 1990 ; Blauert, 1983 ; M. B. Gardner, 1973 ; Hebrank & Wright, 1974). Hebrank & Wright (1974) rapportent que les modifications spectrales ont lieu sur la gamme de fréquence entre 4000 et 16000 Hz et qu'elles ont lieu sur des intervalles différents en fonction de l'élévation. Par exemple, pour une source sonore présentée au niveau de l'axe médian, ils mesurent des trous spectraux entre 4000 et 10000 Hz, et une amplification de l'énergie dans les fréquences au-delà de 13000 Hz. Pour des sources sonores en hauteur (i.e., élévation > 0°), ils rapportent un pic fréquentiel entre 7000 et 9000 Hz et une absorption marquée des fréquences au-delà de 10000 Hz.

Néanmoins, plusieurs études s'accordent à dire que des modifications spectrales dépendantes de l'élévation ont aussi lieu dans les basses fréquences (Algazi et al., 2001a ; Asano et al., 1990 ; M. B. Gardner, 1973). Si Asano et al. (1990) s'accordent avec les résultats de Hebrank et Wright (1974) en rapportant une prédominance des trous spectraux et des amplifications au-delà de 5000 Hz, ils rapportent également des modifications spectrales dans les fréquences inférieures

à 2000 Hz, ce qu'observe aussi Algazi et al. (2001a). Garner (1973) rapportait également la présence de modifications spectrales pouvant être perçues dans la gamme de fréquence entre 700 et 3500 Hz.

De façon générale, lorsqu'une source sonore est diffusée à une basse élévation (i.e., en dessous de la hauteur des oreilles), l'énergie dans les hautes fréquences a tendance à être atténuée. À l'inverse, elle a tendance à être plus préservée lorsque l'élévation de la source sonore est haute. Ces modifications spectrales peuvent être décrites par des fonctions de transfert appelées HRTFs, pour *Head Related Transfer Fonctions* en anglais qui consistent en une paire de fonctions de transfert ($HRTF_G$, $HRTF_D$) reproduisant la signature spectrale d'une position en élévation (i.e., les atténuations et amplifications dans le spectre de fréquence). La **Figure II-11** présente des exemples de modifications spectrales appliquées sur un bruit blanc pour trois élévations différentes (-22.5 °, 0° ou +45°). Alors que le spectre de fréquence d'un bruit blanc généré est plat, des amplifications et atténuations (aussi appelées trous spectraux) marquées apparaissent en fonction de l'élévation à laquelle le signal est simulé.

Figure II-11. Localiser l'élévation d'une source sonore. Modifications spectrales en fonction de la localisation en élévation (+45°, 0° et -22.5°) d'un bruit blanc dont l'azimut et la distance sont fixes (azimut = 0° et distance = 1 m). Les spectres de fréquences du signal gauche stéréophonique (S_G) pour chaque élévation (-22.5° en jaune, 0° en vert, +45° en bleu, et avant spatialisation en noir) sont obtenus après la convolution du signal acoustique d'origine (S) avec les HRIRs correspondants de la base de données CIPIC (Algazi et al., 2001b).



2.2.3. Indices pour la distance

Pour localiser la distance d'une source sonore, le système auditif se repose principalement sur les indices acoustiques que sont l'intensité du signal et des indices de réverbération, mais aussi des indices spectraux et binauraux dans une plus moindre mesure (pour revue, voir Blauert, 1983 ; Zahorik, 2005). Lorsqu'une source sonore émet un son, les ondes sonores s'atténuent au cours de la propagation. Dans un environnement anéchoïque (i.e., limitant les réflexions des ondes sonores), l'amplitude de l'onde sonore s'atténue de façon inversement proportionnelle à la distance. L'atténuation de l'intensité peut être approximée par la fonction suivante (aussi appelée loi du carré inverse) :

$$\frac{I_0}{d^2},$$

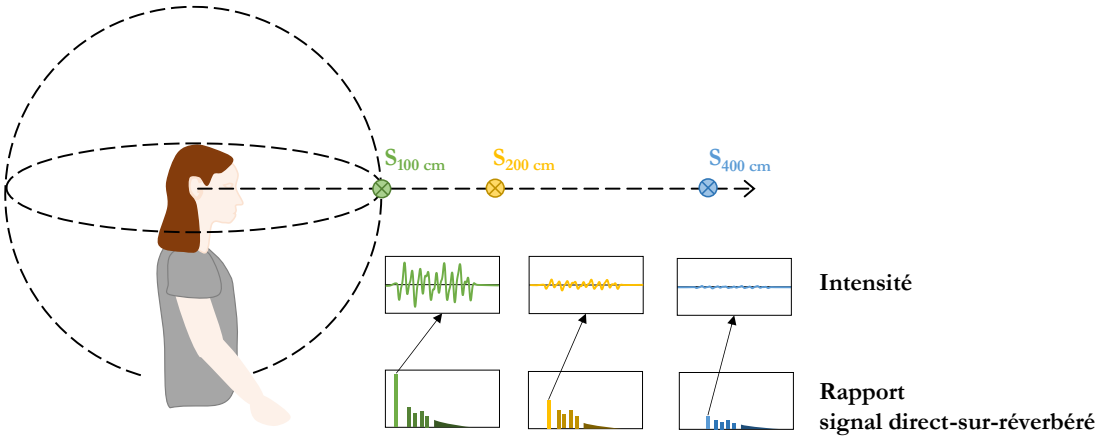
Avec I_0 l'intensité mesurée lorsque la source sonore est à 1 m de l'auditeur, et d la distance actuelle entre la source sonore et l'auditeur (en mètre), **Figure II-12**. Ceci équivaut à une diminution d'environ 6 dB SPL à chaque fois que la distance double. L'amplitude de l'onde s'atténuant, le volume perçu par l'auditeur diminue. L'intensité en elle-même est rarement suffisante pour déterminer la distance d'une source sonore étant donné qu'il s'agit d'une caractéristique dépendant également de la puissance à laquelle le son est diffusé à la source (Blauert, 1983). Ainsi, l'intensité perçue (i.e., la sonie) associée à une source sonore ne dépend pas uniquement de la distance de la source sonore. Néanmoins, Zahorik et Wightman (2001) ont démontré que le système auditif pouvait différencier les changements provenant d'une modification de la distance de la source sonore, de ceux provenant de la puissance à laquelle le son est diffusé à la source.

Lorsqu'un son se propage dans un environnement réverbérant, une onde sonore directe atteint chaque organe récepteur, mais elle est rapidement suivie d'ondes issues des réverbérations. À mesure que la source sonore est éloignée de l'auditeur, le rapport entre le signal direct et le signal réverbéré diminue (**Figure II-12**). Le système auditif a la capacité d'interpréter ces différences d'énergie entre les ondes sonores directes et celles propres à la réverbération pour déterminer la distance d'une source sonore. Ces indices de réverbération peuvent être déterminants pour localiser la distance d'une source sonore, comme le suggère notamment Mershon et Bowers (1979) et Mershon et King (1975) qui mesurent de meilleures performances de localisation dans un environnement réverbérant (i.e., rendant possible la présence d'indices de réverbération) que dans un environnement anéchoïque.

Si l'intensité et les indices de réverbération sont les indices acoustiques majeurs utilisés par le système auditif pour localiser la distance d'une source sonore, des indices spectraux et binauraux peuvent également être utilisés (Blauert, 1983). Par exemple, lorsqu'une source sonore est placée à

une distance éloignée (supérieure à 15 m), les hautes fréquences ont tendance à s'atténuer davantage que les basses fréquences, rendant possible une estimation de la distance si l'on a une connaissance préalable du spectre au niveau de la source sonore (Coleman, 1968). Inversement, pour des sources sonores proches, les effets de filtrages du pavillon et du corps de l'auditeur en général fournissent des indices acoustiques pertinents pour l'estimation de la distance (Brungart et al., 1999). Néanmoins, comme le mentionne Zahorik (2005), les différences intéraurales temporelles (ITD) et en intensité (ILD) étant peu dépendantes de la distance de la source sonore au-delà de 1 m, elles sont majoritairement utilisées lorsque la source sonore est proche de l'auditeur (Brungart et al., 1999 ; Coleman, 1968).

Figure II-12. Localiser la distance d'une source sonore. Atténuation de l'intensité du signal acoustique en fonction de la distance de la source sonore, et schématisation de la diminution du rapport signal direct-sur-réverbéré à mesure que la source sonore est distante de l'auditeur.



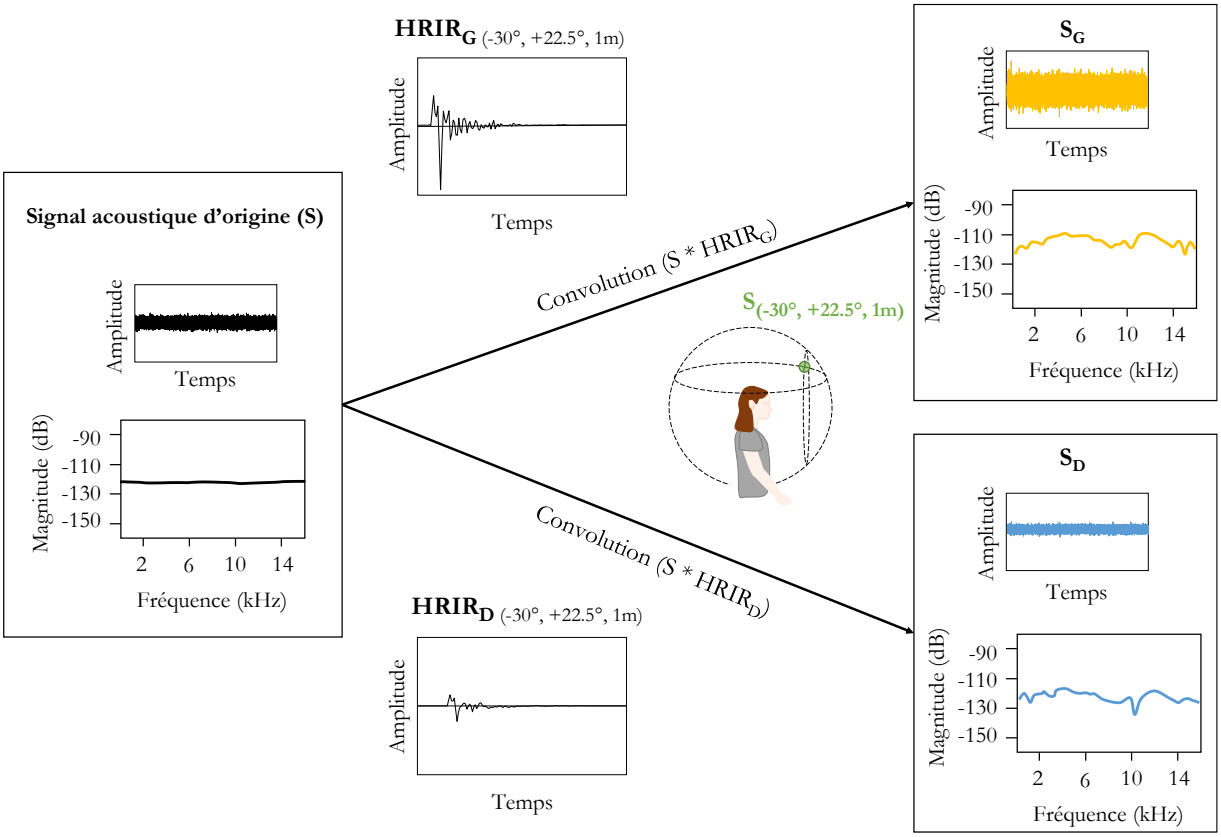
2.2.4. Spatialisation par reproduction des indices acoustiques avec des HRTFs

La spatialisation consiste à simuler une source sonore à une position de l'espace en 3-dimensions en synthétisant un signal stéréophonique intégrant des indices acoustiques spatiaux à partir d'un signal acoustique originellement monophonique. Elle permet la création d'un espace virtuel acoustique (VAS pour *Virtual Acoustic Space* en anglais). Ces indices acoustiques peuvent être intégrés en utilisant les *Head Related Transfer Functions* (HRTFs) qui sont des fonctions transfert caractérisant les modifications que subit un son lors de sa propagation à partir d'une position 3-dimensionnelle dans l'espace. Les HRTFs permettent de transformer un signal acoustique monophonique en un signal acoustique stéréophonique contenant les indices acoustiques tels que les ITDs, ILDs, IPDs, les modifications spectrales et l'atténuation du signal (détaillées dans les sections précédentes II.2.2.1, II.2.2.2 et II.2.2.3). Alors que les HRTFs représentent ces modifications dans le domaine spectral, les Head Related Impulse Responses (HRIRs) représentent ces modifications dans le domaine temporel. L'acquisition des HRTFs et HRIRs repose sur un

enregistrement binaural des signaux acoustiques reçus par chaque oreille lorsqu'un clic est diffusé à plusieurs positions dans l'environnement de propagation (Li & Peissig, 2020).

Le principe de spatialisation repose alors sur la convolution d'un signal acoustique monophonique S avec deux réponses impulsionnelles HRIRs (chacune associée à une oreille). Il en résulte un signal acoustique stéréophonique (**Figure II-13**) simulant la source sonore S à l'endroit correspondant à la position d'enregistrement des HRIRs.

Figure II-13. Spatialisation avec HRIRs. Un bruit blanc monophonique (S) est spatialisé par convolution avec un couple ($HRIR_G$, $HRIR_D$) mesuré à la position (-30°, +22.5°, 1m) issu de la base de données CIPIC (Algazi et al., 2001b). Le signal stéréophonique généré (S_G, S_D) contient les modifications binaurales et spectrales associées à la position 3-dimensionnelle à simuler. Les modifications spectrales sont visibles sur les spectres de fréquence de S_G et S_D alors que les modifications binaurales sont visibles sur les tracés des signaux acoustiques de S_G et S_D.



Lorsque les enregistrements des HRTFs sont effectués auprès d'un auditeur différent de celui qui entendra les sons spatialisés ou auprès d'un mannequin (e.g., mannequin Kemar), les HRTFs sont dites non-individualisées ou génériques. L'utilisation de HRTFs non-individualisées pour la spatialisation présente comme inconvénient de diminuer les performances de localisation en augmentant les erreurs de localisation. Par exemple, elle augmente le risque de confusion entre une source sonore provenant de devant ou de derrière (*front-back confusion* en anglais), d'en haut ou d'en bas, et réduit l'externalisation, c'est-à-dire la perception d'une source sonore externe à la tête de l'auditeur (Best et al., 2020 ; Wenzel et al., 1993). Pour pallier cette potentielle dégradation des

performances de localisation, les HRTFs peuvent être individualisées de plusieurs façons (Xu et al., 2007). Si l'enregistrement des HRTFs peut être effectué sur l'auditeur en question, la procédure est longue et fastidieuse, nécessitant du matériel et des infrastructures spécifiques (Li & Peissig, 2020). L'individualisation des HRTFs peut aussi se faire a posteriori des enregistrements, par exemple en utilisant une méthode de sélection subjective des HRTFs, en moyennant des HRTFs issus d'enregistrements sur plusieurs auditeurs ou en se reposant sur des mesures anthropométriques. Une autre solution pour limiter la dégradation des performances de localisation avec des HRTFs non-individualisées ou génériques est de mettre en place un entraînement pour que l'auditeur s'adapte aux nouveaux indices acoustiques (Berger et al., 2018 ; Mendonça, 2014 ; Stitt et al., 2019).

2.3. Capacités de localisation de sources sonores

La perception spatiale auditive peut être abordée sous l'angle de la perception absolue ou de la perception relative. Alors que la perception absolue consiste en la localisation d'une source sonore unique relativement à l'auditeur, la perception relative consiste en un jugement de la localisation d'une source sonore relativement à une autre source sonore. Les capacités de localisation de sources sonores peuvent être évaluées avec différentes tâches expérimentales et méthodes de réponse, pouvant influencer les performances. Chez des personnes voyantes, Bahu et al. (2016) ont comparé trois méthodes de pointage égocentriques pour indiquer la direction d'une source sonore (tête, main, proximal). Les auteurs ont observé des patterns de réponse similaires entre les méthodes, bien qu'une plus grande précision fût mesurée avec la méthode de pointage dite « proximale » lorsque les sources sonores étaient dans l'hémichamp haut.

En comparant neuf méthodes de réponses pour localiser la direction d'une source sonore chez des personnes non-voyantes, Harber et al. (1993) suggèrent que les meilleures performances étaient obtenues avec les méthodes de pointage utilisant un outil (pointeur) ou sans outil (tête, le buste, l'index). Dans cette section, nous différencierons également le type de sources sonores en fonction de leur nature : réelle, virtuelle avec HRTFs individualisées, ou virtuelle avec HRTFs non-individualisées.

2.3.1. Localiser l'azimut

De manière générale, le système auditif est plus performant pour localiser la position en azimut d'une source sonore présentée dans l'axe médian que latéralement (Blauert, 1983). Pour ce qui est de la perception absolue, ceci a été observé avec des sources sonores réelles (Bahu et al., 2016 ; Makous & Middlebrooks, 1990 ; Oldfield & Parker, 1984 ; Tabry et al., 2013) ou virtuelles avec HRTFs individualisées (Majdak et al., 2010), ou non-individualisées (Parseihian & Katz, 2012 ; Wenzel et al., 1993).

Par exemple, Makous et Middlebrooks, (1990) rapportent une augmentation de l'erreur non signée passant de 1.5 ± 0.4° pour une source sonore réelle placée à une position frontale (azimut de 0°) à 9.7 ± 4.6° pour une source sonore placée à une position latérale (azimut de ±80°). Cette augmentation de l'erreur était associée à une diminution de la précision dans la réponse. Bahu et al. (2016) mesurent également une augmentation de l'erreur non signée en azimut passant de 4° à 10° à mesure de l'augmentation de l'azimut de la source sonore (azimut allant de 0° à ±80°). De façon intéressante, Makous et Middlebrooks (1990) montrent les bénéfices des mouvements de la tête sur les performances de localisation, avec des erreurs inférieures en condition de boucle fermée (*closed-loop* en anglais) comparativement à boucle ouverte (*open-loop*). Effectivement, dans un paradigme de boucle fermée, la possibilité d'effectuer des mouvements de la tête pendant que la source sonore est diffusée permet d'utiliser les modifications des indices acoustiques spatiaux, contrairement à un paradigme de boucle ouverte où la tête est fixe et/ou la durée de présentation de la source sonore est trop courte pour qu'un mouvement de tête soit effectué. Dans leur étude, Makous et Middlebrooks (1990) observent également que l'élévation de la source sonore peut également influencer les performances de localisation de l'azimut, avec une tendance à l'augmentation de l'erreur non signée pour les sources sonores situées en hauteur ou vers le bas.

Dans leur étude, Oldfield et Parker (1984) rapportent une tendance à la surestimation de la latéralité de sources sonores latérales**.** En d'autres termes, les participants avaient tendance à percevoir les sources sonores latérales comme étant plus excentriques qu'en réalité. Ils n'attribuent pas ce biais à un biais moteur inhérent à la méthode de pointage mais suggèrent qu'il s'agirait d'un biais de nature perceptive dû au cône de confusion. Le cône de confusion (voir **Figure II-10**, section II.2.2.2) est notamment à l'origine des confusions devant-derrière (*front-back confusions* en anglais) qui consistent à percevoir une source sonore derrière à la place de devant (et inversement) symétriquement par rapport à l'axe des oreilles. Ce type de confusion est réduit par les indices acoustiques spatiaux dynamiques disponibles dans les paradigmes *closed-loop* (Makous & Middlebrooks, 1990), et pourrait expliquer la surestimation latérale des sources sonores excentriques dans les paradigmes en *open-loop*. Néanmoins, cette surestimation latérale n'est pas systématiquement observée. Par exemple, Tabry et al. (2013) rapportent à l'inverse une tendance à une perception compressée de l'azimut. Notons que les sources sonores dans l'étude de Tabry et al. (2013) étaient localisées uniquement frontalement (entre -90° et +90°) alors que les sources sonores pouvaient être localisées derrière les participants dans les études de Makous et Middlebrooks (1990) ainsi que dans l'étude de Oldfield et Parker (1984). Avec des sources sonores simulées, les capacités de localisation de l'azimut sont également meilleures lorsque la source sonore est située autour de l'axe médian, que ce soit avec des HRTFs individualisées (Majdak et al., 2010) ou non-individualisées (Wenzel et al., 1993). Néanmoins, Wenzel et al., (1993) observent un

nombre plus important des *front-back confusions* pour des sources sonores simulées comparativement à des sources sonores réelles.

Les capacités de localisation relative de l'azimut sont couramment évaluées avec la mesure de l'angle minimum audible (*Minimum audible angle*, MAA en anglais). Le système auditif peut différencier l'azimut de deux sources sonores de basses fréquences (entre 250 et 1000 Hz) lorsqu'elles sont séparées de seulement 1° autour de l'axe médian (Mills, 1958). À mesure que les sources sonores sont situées plus latéralement, les capacités de discrimination diminuent, et le MAA augmente. Par exemple, Mills (1958) montre que le MAA augmente exponentiellement en passant d'une valeur de 1° sur l'axe médian à une valeur d'environ 10° pour des sources sonores localisées autour de 75° en azimut.

Le contenu spectral de la source sonore influence aussi les capacités de localisation de l'azimut. La localisation de l'azimut repose principalement sur l'ITD et l'ILD, mais ces indices sont avant tout présents respectivement pour des fréquences en dessous de 1500 Hz et au-delà de 3000 Hz (Blauert, 1983 ; Risoud et al., 2018). Par conséquent, les capacités de localisation de l'azimut diminuent pour des sources sonores dont le contenu spectral est compris entre 1000 et 3000 Hz (Risoud et al., 2018). L'enveloppe du son peut dans certains cas dégrader les performances de localisation de l'azimut lorsqu'il s'agit de sons à large bande, et améliorer les performances lorsqu'il s'agit de tonalités (i.e., à bandes étroites) (Yost, 2017).

2.3.2. Localiser l'élévation

Tout comme pour l'azimut, les capacités de localisation de l'élévation d'une source sonore semblent meilleures lorsque la source sonore est présentée au niveau de l'axe médian, comparativement à une localisation plus basse ou plus haute. Pour la perception absolue de l'élévation, c'est le cas avec des sources sonores réelles (Bahu et al., 2016 ; Makous & Middlebrooks, 1990 ; Tabry et al., 2013), virtuelles avec HRTFs individualisées (Majdak et al., 2010), et virtuelles avec HRTFs non-individualisées (Wenzel et al., 1993).

Avec des sources sonores réelle, Makous et Middlebrooks (1990) rapportent une augmentation de l'erreur non signée à mesure que la source sonore s'éloigne de l'axe médian (erreur non signée d'environ 3.5 ± 1.0° pour une élévation à +5°, augmentant jusqu'à 10.2 ± 6.7° pour des élévations entre ±25° et ±45°). Bahu et al. (2016) observent également une augmentation de l'erreur non signée (environ 10°) à mesure que l'élévation de la source sonore augmente (élévation passant de 26 à 57°). De plus, ils observent un biais de compression avec une surestimation d'environ 6° pour une élévation de +26° et une sous-estimation d'environ -10° pour une élévation de 57°. Ce biais de compression, qui se manifeste par une surestimation de l'élévation des sources sonores situées dans l'hémichamp bas, et une sous-estimation des sources sonores situées dans le

haut hémichamp, est aussi observé dans l'étude de Tabry et al. (2013), qui rapportent quant à eux des erreurs non signées en élévation d'environ 17°.

Avec des sources sonores virtuelles spatialisées avec des HRTFs individualisées, Majdak et al. (2010) évaluent les capacités de localisation d'un bruit blanc avec des tâches de pointage. Ils rapportent une erreur moyenne de -8.7° pour les hautes élévations et 5.0° pour une élévation à 0°. Wenzel et al. (2013) rapportent quant à eux une augmentation des confusions haut-bas (*up-down confusions* en anglais) avec des sources sonores simulées avec des HRTFs non-individualisées comparativement à des sources sonores réelles. De plus, avec des HRTFs non-individualisées, de plus grandes erreurs angulaires en élévation sont souvent mesurées, comme dans les études de Mendonça et al. (2013) et Geronazzo et al. (2018) qui rapportent des erreurs non signées de respectivement environ 29.3° et entre 15.58° et 33.75°. Ainsi, que ce soit pour des sources sonores réelles ou virtuelles, ces études montrent une tendance à l'augmentation de l'erreur à mesure que la source sonore s'éloigne de l'axe médian.

En termes de localisation relative de l'élévation, le plus petit MAA a été suggéré lorsque la source sonore est présentée en face de l'auditeur, avec un MAA entre 3.2° et 4.3° (Wettschureck, 1973). Le MAA augmentait jusqu'à environ 10° avec l'augmentation de l'élévation de la source sonore (entre 60° et 90° d'élévation).

La localisation de l'élévation reposant principalement sur l'analyse d'indices spectraux, la complexité spectrale et la largeur de bande influencent les capacités de perception de l'élévation. Il en découle des difficultés à localiser l'élévation d'une source sonore à faible largeur de bande, ou ne contenant pas de fréquences au-delà de 3000 Hz (Algazi et al., 2001a ; Asano et al., 1990 ; Blauert, 1983 ; M. B. Gardner, 1973 ; Hebrank & Wright, 1974).

2.3.3. Localiser la distance

De façon générale, les capacités d'estimation de la distance d'une source sonore diminuent avec l'augmentation de la distance de celle-ci (Zahorik, 2005). Les capacités de localisation de la distance d'une source sonore sont très souvent caractérisées par un biais de compression (pour revue, voir Kolarik et al., 2016 ; Zahorik, 2005). Ce biais de compression a été rapporté pour des sources sonores réelles (Loomis et al., 1998 ; Parseihian et al., 2014), virtuelles avec HRTFs individualisées (Bronkhorst & Houtgast, 1999 ; Kopčo & Shinn-Cunningham, 2011 ; Zahorik, 2002), virtuelles avec HRTFs non-individualisées (Kolarik et al., 2020 ; Martin et al., 2021). Ce biais de compression se traduit par une tendance à surestimer la distance de sources sonores proches et à sous-estimer la distance de sources sonores éloignées. La perception de la distance peut alors être approximée par une fonction puissance $d' = k \times d^a$, avec d' la distance perçue estimée, d la distance physique de la source sonore, et avec k représentant la distance véridique (la distance pour laquelle

la surestimation passe à la sous-estimation des distances) et *a* représentent l'intensité du biais de compression (si $a < 1$). En analysant les fonctions puissances rapportées dans 21 études, Zahorik (2005) rapporte une valeur moyenne *a* d'environ 0.54 et une valeur moyenne de *k* d'environ 1.32. Alors que la valeur *a* inférieure à 1 met en évidence un biais de compression, la valeur de *k* retranscrit la distance seuil à partir de laquelle la distance des sources sonores n'est plus surestimée mais sous-estimée (i.e., distance véridique). Zahorik (2002) avait également montré la persistance du biais de compression avec des valeurs *a* toutes inférieures à 1 et allant de 0.15 à 0.70, mais rapportait à l'inverse une plus grande variabilité dans les constantes *k* qui allaient de 0.39 à 4.04 en fonction des études. Malgré la variabilité des types de sons à localiser, des intervalles de distances testés et des méthodes de réponse utilisées dans ces études, la perception de la distance de sources sonores est presque toujours approximée par une fonction psychophysique puissance dont la courbe démontre un biais de compression. Dans Martin et al. (2021), où des sources sonores étaient simulées entre 1 et 7 m avec des HRTFs non-individualisées, un biais de compression de la distance a également été mesuré. Par contre, en utilisant également des HRTFs non-individualisées mais en simulant des sources sonores très proches (entre 33 et 85 cm), les participants de Parseihian et al. (2014) ne parvenaient pas à distinguer différentes distances de la source sonore.

Ces difficultés de perception de la distance peuvent être liées à un déficit d'externalisation lorsque des sources sonores virtuelles sont utilisées (Leclère et al., 2019 ; Mendonça et al., 2013, voir Best et al., 2020 pour revue). Un déficit d'externalisation est intrinsèquement lié à une non-perception de la distance, puisque la source sonore a tendance à être perçue à l'intérieur de la tête de l'auditeur (Best et al., 2020). L'individualisation des HRTFs n'améliorant pas systématiquement l'externalisation (e.g., Leclère et al., 2019), il semblerait que les indices acoustiques spectraux spécifiques à l'auditeur ne soient pas indispensables à l'externalisation. Par contre, les indices de réverbérations seraient critiques pour une perception externalisée d'une source sonore virtuelle (Best et al., 2020), notamment en présence d'indices binauraux (Leclère et al., 2019). La cohérence entre les indices de réverbération contenus dans le signal de la source sonore virtuelle et la représentation interne qu'a l'auditeur de l'environnement de propagation influence aussi l'externalisation (Klein et al., 2017).

En termes de perception relative de la distance, les capacités de discrimination de deux distances sont de l'ordre de 10 % (Zahorik, 2005), mais de meilleures capacités ont déjà été observées. Par exemple, pour des sources sonores réelles placées à ± 1 et ± 2 m de l'auditeur, Ashmead et al. (1990) rapportent des capacités de discrimination de l'ordre de 5.73 et 5.91 %, respectivement.

2.3.4. Localiser une source sonore dans une scène auditive complexe

Au quotidien, les scènes auditives perçues sont composées d'un ensemble d'évènements sonores qui se succèdent et s'entremêlent. La représentation de la scène auditive repose sur la séparation et le regroupement de régularités spectro-temporelles contenues dans les signaux acoustiques (Bregman, 1990 ; Lemaitre et al., 2018). Pour localiser une source sonore au sein d'une scène auditive, le système auditif doit pouvoir distinguer les différentes sources sonores la composant pour séparer la source sonore à localiser parmi les autres sources sonores non-pertinentes. Il s'agit typiquement d'un contexte de *cocktail party* (Bronkhorst, 2000 ; Cherry, 1953), qui traduit la capacité du système auditif à séparer différents événements sonores dans une scène auditive complexe (e.g., une fête), identifier ces événements sonores, porter son attention sur un événement sonore d'intérêt à un moment (i.e., une conversation), puis sur un autre, et ainsi de suite.

De façon générale, les capacités de localisation d'une source sonore diminuent à mesure que le nombre de sources sonores augmente, ce qui a été observé avec des sources sonores réelles (Brungart et al., 2005 ; Brungart et al., 2014 ; Zhong & Yost, 2017) mais aussi avec des sources sonores virtuelles (Feierabend et al., 2019 ; Kawashima & Sato, 2015). Il a été montré que les capacités à identifier le nombre de sources sonores et à les localiser chutent drastiquement au-delà de 5 sources sonores à large bande présentées simultanément (Brungart et al., 2005 ; Zhong & Yost, 2017), et au-delà de 3 sources sonores à bande étroite (Zhong & Yost, 2017). Par exemple, Brungart et al. (2005) ont mesuré les capacités de localisation d'une source sonore présentée parmi d'autres sources sonores (jusqu'à 14 sources simultanées), réparties à différentes élévations et azimuts. Ils ont mesuré des erreurs non signées augmentant progressivement de 5° lorsqu'une unique source sonore était à localiser, jusqu'à 30° lorsqu'elle était diffusée parmi 4 autres. Au-delà de 5 sources sonores, ils ont observé une augmentation drastique de l'erreur atteignant 80° lorsque la source sonore était à localiser parmi 13 autres sources sonores. Tout comme dans l'étude de Makous et Middlebrooks (1990), Brungart et al. (2005) ont montré l'avantage d'un paradigme *closed-loop* comparativement à un paradigme *open-loop* qui permet une utilisation d'indices acoustiques dynamiques pour localiser une source sonore, qu'elle se trouve ou non dans une scène auditive complexe. Néanmoins, les auteurs suggéraient une limite de ces indices acoustiques dynamiques au-delà de 5 sources sonores simultanées allant jusqu'à la disparition de l'effet facilitateur des mouvements de tête (*closed-loop*) sur les performances de localisation.

L'arrangement spatial entre les différentes sources sonores influence les capacités de ségrégation et de localisation des sources sonores (Kwak et al., 2020 pour revue). De façon générale, les capacités à ségréguer des sources sonores dans une scène auditive dépendent du nombre de sources sonores et sont facilitées par la séparation spatiale des sources sonores. Ce phénomène est

appelé *spatial release from masking*. Ceci a été suggéré par plusieurs études, comme celles de Kawashima et Sato (2015) et Zhong et Yost (2017) qui ont montré qu'augmenter la distance séparant les sources sonores réduisait l'altération des capacités de localisation dans une scène auditive complexe.

En présence d'un bruit de fond, les capacités à localiser une source sonore diminuent avec l'augmentation du volume du bruit de fond relativement à celui de la source sonore (i.e., diminution du rapport signal-sur-bruit) (Kerber & Seeber, 2012 ; Lorenzi et al., 1999). Pour séparer deux sources sonores ou localiser une source sonore parmi un bruit de fond, la localisation en azimut et en élévation des différentes sources sonores ainsi que leur arrangement spatial influencent les capacités de séparation et de localisation. Par exemple, les capacités à séparer deux sources sonores présentées frontalement et séparées en azimut sont meilleures que lorsqu'elles sont présentées latéralement, avec des sources sonores virtuelles à large bande (Best et al., 2004) et des tonalités (Perrott, 1984). En présence d'un bruit de fond dont l'intensité varie, Lorenzi et al. (1999) montrent également que l'altération des performances de localisation provoquée par la diminution du rapport signal-sur-bruit est plus importante lorsque le bruit de fond est présenté latéralement que frontalement. Par contre, pour deux sources sonores localisées latéralement, Best et al. (2004) suggèrent qu'il est plus difficile de les séparer lorsqu'elles sont séparées en azimut plutôt qu'en élévation.

2.3.5. Capacités de localisation de sources sonores chez les personnes non-voyantes

En ce qui concerne la perception spatiale, l'expérience visuelle participe à l'acquisition d'une représentation interne de l'espace physique. Dans le cas des personnes non-voyantes congénitales ou précoces, l'absence de cette expérience visuelle modifie la calibration de l'espace. Les capacités de localisation de sources sonores chez les personnes non-voyantes sont caractérisées par une variabilité inter-individuelle mettant parfois en évidence des performances supranormales, et à l'inverse parfois déficitaires (pour revue, voir Kolarik et al., 2016 ; Voss et al., 2010).

Plusieurs études tendent à mettre en évidence des capacités de perception auditive spatiale supérieures chez les non-voyants pour l'azimut (Doucet et al., 2005 ; Lessard et al., 1998 ; Voss et al., 2015), l'élévation (Voss et al., 2015) et la distance (Kolarik et al., 2013). Ces capacités supranormales émergeraient de mécanismes de compensation du fait de la privation sensorielle visuelle. Pour localiser l'azimut d'une source sonore, Doucet et al. (2005) et Voss et al. (2015) ont montré une variabilité inter-individuelle dans les performances de localisation chez les non-voyants, certains ayant des capacités supérieures aux voyants, et d'autres comparables aux voyants. Les non-voyants ayant de meilleures performances semblaient utiliser les indices spectraux à meilleur

escient, suggérant des capacités de perception auditives spatiales supérieures chez les personnes non-voyantes. En mesurant les performances de discrimination de la distance de sources sonores simulées avec des HRTFs non-individualisées, Kolarik et al. (2013) ont observé des capacités supérieures chez les non-voyants que chez les voyants. Ces résultats suggèrent que les non-voyants utilisent plus efficacement les indices acoustiques tels que l'intensité et les réverbérations pour la perception auditive de la distance.

À l'inverse, d'autres études mettent en évidence des capacités de localisation auditive réduites chez les personnes non-voyantes pour l'azimut, l'élévation (Lewald, 2002 ; Voss et al., 2015) et la distance (Kolarik et al., 2020). Par exemple, Lewald (2002) montre des performances de localisation de l'élévation inférieures chez les personnes non-voyantes que chez les personnes voyantes. Pour la localisation de la distance, les résultats de Kolarik et al. (2020) suggèrent que la surestimation des distances proches est plus marquée chez les personnes non-voyantes que voyantes. Également, Voss et al. (2015) observent des performances de localisation de l'élévation réduites chez certains participants non-voyants. Les participants montrant un déficit dans les capacités de localisation de l'élévation étaient également les participants qui avaient les meilleures performances de localisation pour la dimension de l'azimut. En d'autres termes, ces résultats suggèrent que les mécanismes de compensation auditive chez les personnes non-voyantes ont des limites.

Dans le contexte de l'analyse de scènes auditives complexes, par exemple, dans une scène auditive reproduisant une situation de *cocktail party*, Feierabend et al. (2019) observent que la présence de plusieurs sources sonores (i.e., diminution du rapport signal-sur-bruit) entraînait une dégradation similaire des performances de localisation dans l'azimut entre les voyants et les non-voyants. Pourtant, dans une autre étude, Zwiers et al. (2001) ont montré que les capacités des personnes non-voyantes à localiser l'élévation d'une source sonore présentée avec un bruit de fond étaient inférieures à celle des voyants.

Ces résultats à première vue hétérogènes mettent en évidence que des mécanismes compensatoires existent, mais qu'ils ne permettent pas toujours de compenser les déficits de calibration de la représentation spatiale causés par la privation visuelle.

2.4. Interactions audio-visuelles spatiales

2.4.1. Association entre hauteur tonale et hauteur spatiale : correspondance cross-modale et valence spatiale

Les correspondances cross-modales consistent en une tendance à associer une dimension de stimulus d'une modalité (e.g., visuelle) avec la dimension d'un stimulus d'une autre modalité

(e.g., auditive) (pour revue, voir Spence, 2011). Parmi les correspondances cross-modales audiovisuelles répertoriées, on retrouve celle entre la hauteur tonale d'un son et la hauteur spatiale d'un stimulus visuel. Elle se manifeste par une tendance à associer des sons aigus à des stimuli visuels localisés en hauteur, et inversement des sons graves à des stimuli visuels localisés plus bas. D'un point de vue comportemental, elle entraîne, par exemple, des temps de réaction plus courts dans une tâche audio-visuelle de Go/No-Go lorsque les stimuli visuels et auditifs sont congruents, c'est-à-dire lorsque la hauteur du son est plus élevée et que l'emplacement visuel est plus élevé, et lorsque la hauteur du son est plus faible et que l'emplacement visuel est plus bas (Miller, 1991). Dans une tâche de discrimination de position verticale d'un stimulus visuel, la rapidité et la précision sont plus élevées lorsque la hauteur tonale d'un son diffusé est congruente avec la hauteur du stimulus visuel (Evans & Treisman, 2011). La correspondance audiovisuelle entre hauteur tonale et hauteur spatiale entraîne donc des facilitations de traitement d'informations visuelles lorsqu'une tonalité congruente en fréquence est présentée simultanément.

L'association entre hauteur tonale d'un son et hauteur spatiale a aussi été mise en évidence dans des tâches auditives de localisation (Pedley & Harper, 1959 ; Pratt, 1930) ou de comparaison de hauteur tonale et de timbre (Rusconi et al., 2006). Dans les tâches de localisation utilisées par Pratt (1930) et Pedley et Harper (1959), les participants avaient tendance à localiser la source sonore à une position plus haute lorsqu'il s'agissait d'une tonalité aiguë. Surtout, dans l'étude de Pedley et Harper (1959), lorsque la source sonore à localiser était plus grave que les autres tonalités pouvant être diffusées, les participants la localisaient à une position plus basse que lorsque cette même tonalité était diffusée avec des tonalités plus graves. Leur étude suggère alors une correspondance relative plutôt qu'absolue. Dans l'étude de Rusconi et al. (2006), l'association entre hauteur tonale et hauteur spatiale se manifeste par des temps de réponse plus rapides pour des sons aigus lorsque le bouton de réponse à utiliser était celui le plus haut (et pour des sons graves lorsque le bouton de réponse à utiliser était celui le plus bas). Ces études semblent démontrer que nous avons tendance à nous représenter spatialement l'échelle de fréquence, notamment sur un axe vertical.

2.4.2. Influence auditive sur des évènements visuels spatiaux

Dans un contexte multisensoriel, les informations présentées dans une modalité sensorielle peuvent influencer la perception spatiale d'informations d'une autre modalité sensorielle. C'est typiquement le cas de l'effet ventriloquisme (Howard, 1966). Dans cet effet, des informations auditives (paroles) sont localisées comme provenant d'un objet visuel situé à un autre endroit que la source sonore (la marionnette). Un autre effet audio-visuel est celui d'induction de rebond (*audiovisual bounce-inducing effect* en anglais) (Sekuler et al., 1997). Cet effet se manifeste dans une situation où deux disques identiques se déplacent sur un écran et qu'un son est diffusé lorsqu'ils se

croisent. Dans cette situation, deux évènements spatiaux visuels peuvent être perçus : soit les disques se traversent (i.e., se croisent), soit ils rebondissent l'un contre l'autre et changent de trajectoire. Dans une situation uni-sensorielle visuelle, les disques sont le plus souvent perçus comme se croisant. Mais lorsqu'un son est présenté au même moment que les disques se croisent, la proportion de rebonds perçus augmente. Cet effet est amplifié lorsque l'amplitude de l'enveloppe du stimulus auditif diffusé est décroissante, plutôt que croissante (Grassi & Casco, 2009). Une amplitude d'enveloppe décroissante se rapproche de celle induite par un impact (e.g., deux boules de pétanque entrant en contact). Ainsi, l'amplitude de l'enveloppe sonore peut influencer la perception d'événements spatiaux visuels.

2.4.3. Influence auditive sur la perception spatiale chez les non-voyants

Si des interactions audio-visuelles telles que la correspondance cross-modale entre la hauteur tonale et l'élévation sont observées chez les personnes voyantes, leur présence chez les personnes non-voyantes est discutable.

Pour étudier l'existence d'une correspondance cross-modale entre la hauteur tonale et l'élévation spatiale chez les non-voyants, Deroy et al. (2016) se sont intéressés à la correspondance cross-modale analogique audio-tactile entre la hauteur tonale et la direction d'une stimulation tactile (croissante ou décroissante spatialement). En utilisant une tâche d'association implicite dans laquelle les participants devaient catégoriser le sens d'une stimulation auditive (hauteur tonale croissante ou décroissante) ou tactile (dirigée vers le haut ou le bas), Deroy et al. (2016) ont mesuré des temps de réaction plus rapides pour juger la direction de la stimulation tactile lorsqu'il y avait une congruence que dans la condition incongruente chez les personnes voyantes. Néanmoins, l'effet de congruence n'a pas été mesuré chez les personnes non-voyantes, suggérant que la correspondance audio-tactile entre la direction de la hauteur tonale et celle d'une stimulation tactile ne serait pas présente chez les personnes non-voyantes. Également, l'effet « *Bouba-Kiki* » consistant en une association entre des mots (*Bouba* et *Kiki*) et la forme d'objets (arrondis ou anguleux), a été observé chez des personnes non-voyantes tardives seulement de façon réduite, et n'a pas été détecté chez des non-voyants précoces (Fryer et al., 2014). Si ces deux études suggèrent la nécessité d'une expérience visuelle pour que certaines correspondances cross-modales émergent, l'étude de Hamilton-Fletcher et al. (2020) tend pourtant à démontrer l'existence d'une correspondance entre hauteur tonale et élévation chez les non-voyants. Dans cette étude, l'association entre la hauteur tonale et l'élévation chez les non-voyants a été étudiée en proposant une tâche d'association implicite auditive entre la hauteur tonale d'un son et un mot indiquant une hauteur spatiale (« haut » ou « bas »). En analysant les temps de réaction et les erreurs à la tâche de classification, les résultats ont mis en évidence un effet de congruence chez les non-voyants. Le peu d'études relatives à cette

question a pour conséquence que l'existence de la correspondance cross-modale entre la hauteur tonale et l'élévation peut difficilement être affirmée chez les personnes non-voyantes.

2.5. Synthèse

- La perception auditive est multidimensionnelle. Un son est un percept caractérisé par plusieurs dimensions telles que sa sonie, sa hauteur tonale et son timbre. Ces attributs sont issus de la modulation de dimensions physiques du signal acoustique telles que l'amplitude, la fréquence et l'enveloppe.

- La perception spatiale auditive 3-dimensionnelle repose sur des indices spatiaux acoustiques binauraux (intensité, temporels) et des indices monauraux (spectraux, d'intensité, temporels) qui sont utilisés différemment en fonction de la dimension spatiale (azimut, élévation, distance). Ces indices spatiaux acoustiques peuvent être reproduits dans un signal acoustique avec des HRTFs pour simuler la spatialisation d'une source sonore.

- Les capacités de perception spatiale d'une source sonore présentent des limites propres à l'azimut, l'élévation et la distance. Ces limites dépendent notamment des caractéristiques de la source sonore et de la complexité de la scène auditive à analyser.

- Les capacités de perception auditives et spatiales de la population non-voyante sont marquées par une variabilité inter-individuelle démontrant aussi bien des capacités supranormales que l'inverse.

- L'existence d'interactions audio-visuelles spatiales démontre que certaines dimensions du son (e.g., hauteur tonale, enveloppe) peuvent influencer la perception spatiale.

Dans le contexte du développement d'un DSS vision-vers-audition pour l'aide à la locomotion et à la localisation d'obstacles, la multidimensionnalité du son rend possible l'intégration de nombreux indices acoustiques. La perception spatiale auditive repose sur des indices acoustiques spatiaux reproductibles en utilisant des HRTFs pour simuler le placement d'une source sonore dans un espace en 3 dimensions. Au-delà de ces indices acoustiques spatiaux, la perception spatiale peut être influencée par d'autres dimensions du son. La section suivante présente les indices acoustiques utilisés dans les DSSs existants et les capacités de perception spatiale qui sont associées.

---

3. Perception spatiale avec un dispositif de substitution sensorielle vision-vers-audition

3.1. Indices acoustiques utilisés dans les dispositifs existants

Les éléments graphiques convertis ainsi que les indices acoustiques utilisés par les dispositifs de substitution sensorielle (DSSs) vision-vers-audition sont divers. Elli et al. (2014) et Hamilton-Fletcher et al. (2016b) soulignaient dans leurs travaux l'importance de l'intuitivité des schémas d'encodage qui doivent être retenus. En effet, la difficulté des utilisateurs pour apprendre à interpréter ces nouvelles informations auditives figure parmi les raisons qui font que ces dispositifs ne sont encore que peu adoptés. Le **Tableau II-1Autres dimensions** synthétise les DSSs vision-vers-audition existants en présentant les dimensions visuelles transmises et les indices acoustiques utilisés pour les transmettre dans le paysage sonore. Parmi les éléments graphiques transmis, on retrouve des dimensions visuo-spatiales ainsi que d'autres dimensions visuelles (e.g., couleur, luminosité). Notons que les premiers DSSs transmettaient principalement des informations visuelles 2-dimensionnelles, se limitant à la position horizontale et verticale d'une information visuelle sur une image. Avec le développement des caméras de profondeur, de plus en plus de DSSs encodant également la distance se sont développés. Certains DSSs, dont SoundSight (Hamilton-Fletcher et al., 2022) proposent plusieurs options d'encodage pour différentes dimensions qui sont modulables en fonction des préférences et capacités propres à l'utilisateur.

3.1.1. Axe horizontal (équivalent azimut)

Parmi les DSSs existants, on peut distinguer trois types d'indices acoustiques pour encoder l'axe horizontal de l'image : le scanning temporel, la spatialisation, et la fréquence.

Le scanning temporel consiste à parcourir l'image, couramment de gauche à droite, pour transmettre auditivement des éléments graphiques de chaque colonne l'une après l'autre. Pour une image donnée, le paysage sonore associé est habituellement transmis sur une durée d'une à deux secondes. Cet encodage est utilisé par les dispositifs the vOICe (Meijer, 1992), celui de Cronly-Dillon et al. (1999) et EyeMusic (Abboud et al., 2014). Le dispositif SoundSight proposé par Hamilton-Fletcher et al. (2022) propose également cet encodage parmi les options pour l'axe horizontal. Un inconvénient notable de cet encodage est sa faible résolution temporelle. Par exemple, Proulx et al. (2008) évoquent que l'utilisation du dispositif the vOICe nécessite d'adapter la vitesse des mouvements de tête (i.e., de la caméra) à la faible résolution temporelle du dispositif (entre 1 et 2 secondes) afin de ne pas manquer des informations importantes présentes dans l'image acquise par la caméra.

Tableau II-1. Synthèse de schémas d'encodage utilisés dans des dispositifs de substitution (DSS) vision-vers-audition existants. Les éléments graphiques convertis sont présentés séparément pour les dimensions spatiales (azimut, élévation, distance) et pour d'autres (couleur, luminosité, température). Pour chaque élément graphique transmis par un DSS, les indices acoustiques utilisés pour le convertir sont fournis. (a) HRTFs non-individualisées de la base de données CIPIC (Algazi et al.,2001b), (b) HRTFs non-individualisées, enregistrées par l'équipe, (c) HRTFs non-individualisées de OpenAL, (d) Logiciel audio Reaper, (e) HRTFs non-individualisées de la base de données MIT KEMAR (Gardner & Martin, 1994), (f) HRTFs non-individualisées de la classe SoundStream de la librairie SFML.

Références **Nom du DSS**	**Dimension spatiale**			**Autres dimensions**
	Azimut (Horizontal)	**Elévation (Vertical)**	**Distance**	
Abboud et al. (2014) **EyeMusic**	Scanning temporel stéréophonique	Fréquence [65 Hz, 1760 Hz]	.	Couleur : Timbre
Ambard et al. (2015)	Stéréophonie	Fréquence [250, 2500 Hz]	.	Luminosité : Volume
Bizon-Angov et al. (2021) **Colorphone**	Stéréophonie	.	Scanning temporel	Couleur : Fréquence {250 Hz, 500 Hz, 1000 Hz}
Capelle et al. (1998) **PSVA**	Fréquence [50 Hz, 12526 Hz]	Fréquence [50 Hz, 12526 Hz]	.	.
Commère et al. (2020) **See Differently**	HRTFs (a)	Timbre	.	.
Cronly-Dillon et al. (1999)	Scanning temporel	Fréquence	.	.

Tableau II.1. (suite)

Références **Nom du DSS**	**Dimension spatiale**			**Autres dimensions**
	Azimut (Horizontal)	**Elévation (Vertical)**	**Distance**	
Gonzalez-Mora et al. (1999)	HRTFs (b)	HRTFs (b)	HRTFs (b)	Luminosité : Volume
Hamilton-Fletcher et al. (2016a) **Synaestheatre**	HRTFs (d)	Fréquence [110 Hz, 880 Hz]	Volume	Couleur : Timbre
Hamilton-Fletcher et al. (2022) **SoundSight**	{Scanning temporel, HRTFs}	{Scanning temporel, Fréquence, timbre}	{Volume, Scanning temporel, réverbération}	Couleur : {Timbre, Volume} Température : {Timbre, Volume} Luminosité : Volume
Hanneton et al. (2010) **The Vibe**	Stéréophonie	Fréquence	Luminosité : Volume	.
Meijer (1992) **The vOICe**	Scanning temporel (stéréophonique)	Fréquence [500 Hz, 5000 Hz]	.	Luminosité : Volume
Mhaish et al. (2016)	HRTFs (e)	HRTFs (e)	Volume	.
Neugebauer et al. (2020)	.	Fréquence	Volume	.

Tableau II.1. (suite)

Références **Nom du DSS**	**Dimension spatiale**			**Autres dimensions**
	Azimut (Horizontal)	**Elévation (Vertical)**	**Distance**	
Paré et al. (2021) **GSSD**	Stéréophonie	.	Taux de répétition d' un bip, Fréquence et Volume	.
Pourghaemi et al. (2018)	.	Scanning	.	Largeur de l' objet : Fréquence
Ribeiro et al. (2012)	HRTFs (a)	HRTFs (a)	Volume et Réverbération	.
Richardson et al. (2019) **Synaestheatre**	HRTFs (d)	HRTFs (d)	Volume	.
Spagnol et al. (2017)	HRTFs (e)	*Rising factor*	Volume	.
Stoll et al. (2015) **MeloSee**	Stéréophonie	Fréquence [261.626 Hz, 523.251 Hz]	Volume	.
Ton et al. (2018) **LASS**	HRTFs (f)	.	Fréquence [170 Hz, 650 Hz]	.

Une deuxième méthode pour encoder auditivement l'axe horizontal de l'image repose sur la spatialisation. On distinguera les schémas d'encodage utilisant la stéréophonie de ceux utilisant les HRTFs, bien que les deux méthodes impliquent des écouteurs stéréophoniques. Les dispositifs utilisant la stéréophonie modulent l'intensité relative de chaque canal de diffusion (gauche et droit) et fonction de la position sur l'axe horizontal. La stéréophonie est utilisée par le dispositif Colorphone (Bizoń-Angov et al., 2021), GSSD (Paré et al., 2021), MeloSee (Stoll et al., 2015), the Vibe (Hanneton et al., 2010), et le dispositif proposé par Ambard et al. (2015). Notons que la spatialisation est parfois couplée avec le scanning temporel. C'est le cas de the vOICe et EyeMusic qui couplent le scanning temporel avec la stéréophonie. Les dispositifs utilisant la spatialisation avec HRTFs pour l'axe horizontal spatialisent un signal acoustique à partir de HRTFs. On retrouve le dispositif de Gonzalez-Mora et al. (2006), celui de Ribeiro et al. (2012), See Differently (Commère et al., 2020), SoundSight (Hamilton-Fletcher et al., 2022), celui de Spagnol et al. (2017), Synaestheatre (Hamilton-Fletcher et al., 2016a ; Richardson et al., 2019), et celui proposé par Mhaish et al. (2016). Ils utilisent couramment des HRTFs mis à disposition dans des bases de données, donc non-individualisées. Par exemple, dans le dispositif Synaestheatre, la base de données CIPIC (Algazi et al., 2001b) est utilisée pour spatialiser une tonalité (Hamilton-Fletcher et al., 2016a) ou un bruit blanc (Richardson et al., 2019). L'utilisation de schémas d'encodage reposant sur la spatialisation s'inscrit dans une approche de mimétisme de la modalité de substitution (i.e., ici auditive).

Enfin, le dispositif PSVA (Capelle et al., 1998) encode l'axe horizontal avec la fréquence du son. Pour une ligne de pixels de l'image, la fréquence du son associé à chaque pixel augmente de gauche à droite. Notons que le dispositif PSVA utilise également la fréquence pour encoder l'axe vertical de l'image.

3.1.2. Axe vertical (équivalent élévation)

Pour encoder l'axe vertical de l'image, plusieurs méthodes peuvent être distinguées : la spatialisation, la fréquence, le timbre et le scanning temporel.

Les DSSs utilisant la spatialisation pour encoder l'élévation utilisent des HRTFs non-individualisées (Gonzalez-Mora et al., 2006 ; Ribeiro et al., 2012 ; Richardson et al., 2019). Ces dispositifs se placent dans une approche de mimétisme de la modalité auditive. Néanmoins, afin que les modifications spectrales propres aux indices acoustiques pour l'élévation s'appliquent, ces dispositifs spatialisent des bruits blancs.

Une autre catégorie de DSSs encode l'axe vertical en modulant la fréquence du son (Abboud et al., 2014 ; Ambard et al., 2015 ; Capelle et al., 1998 ; Cronly-Dillon et al., 1999 ; Hamilton-Fletcher et al., 2022 ; Hamilton-Fletcher et al., 2016a ; Hanneton et al., 2010 ; Meijer,

1992 ; Neugebauer et al., 2020 ; Stoll et al., 2015). Si ces dispositifs utilisent des intervalles de fréquences différents et une résolution différente, tous s'accordent à associer les fréquences les plus basses (graves) à une position verticale basse sur l'image, et les fréquences les plus hautes (aigues) aux pixels les plus hauts. Si la modulation de la fréquence peut être qualifiée d'arbitraire, Brown et al. (2015) rappellent qu'elle ne l'est pas totalement puisqu'elle est employée du fait de l'association cross-modale entre hauteur tonale et hauteur spatiale (détails dans la section II.2.4.1). La logique derrière ces dispositifs consiste à dire que si l'indice acoustique (i.e., ici la fréquence) peut faciliter la perception spatiale, l'apprentissage de l'utilisation du DSS sera lui aussi facilité (Kristjánsson et al., 2016). Expérimentalement, l'effet de facilitation de schémas d'encodage utilisant la modulation de la fréquence pour l'axe vertical a été suggéré par Stiles et Shimojo (2015) avec le dispositif the vOICe, et Buchs et al. (2021) avec le dispositif EyeMusic dans des tâches de reconnaissance de formes. Dans ces deux études, les performances étaient supérieures au seuil de l'aléatoire avant tout entraînement.

D'autres dispositifs utilisent des indices acoustiques plus arbitraires comme le timbre (Commère et al., 2020 ; Hamilton-Fletcher et al., 2022), le scanning temporel (Hamilton-Fletcher et al., 2022 ; Pourghaemi et al., 2018) ou un *rising factor* (Spagnol et al., 2017). L'encodage par la modulation du timbre repose sur l'association entre une position verticale de l'image et un instrument de musique ou un type de son (e.g., oiseau). Cet encodage présente l'inconvénient d'être limité à peu de positions verticales puisque qu'il pourrait surcharger la mémoire de travail. Le DSS proposé par Spagnol et al. (2017) utilise des modélisations de sons de bulles dans l'eau. L'encodage sonore de l'axe vertical s'inspire du bruit des bulles qui remontent à la surface, ce qu'ils modélisent avec ce qu'ils appellent un *rising factor*, qui augmente à mesure que la position verticale du pixel augmente.

3.1.3. Distance

Avec les progrès technologiques concernant les caméras et les progrès en traitement d'image, de plus en plus de DSSs cherchent à transmettre des informations sur la distance des objets. La distance est une dimension spatiale présentant un fort intérêt dans le cadre d'un déplacement pédestre puisqu'elle contribue fortement à éviter des obstacles. Pour encoder la distance, les dispositifs existants utilisent des indices de réverbération, la modulation du volume, de la fréquence, ou des indices temporels telle que la vitesse de répétition d'un son, la durée d'un son, ou le scanning temporel.

La modulation du volume et l'incorporation d'indices de réverbération s'inscrivent dans une approche de mimétisme des mécanismes de localisation de la distance de sources sonores (section II.2.2.3). La modulation du volume a l'avantage d'être facilement implémentable. C'est

pour cela qu'elle est majoritairement utilisée (Hamilton-Fletcher et al., 2016a ; Hamilton-Fletcher et al., 2022 ; Neugebauer et al., 2020 ; Paré et al., 2021 ; Ribeiro et al., 2012 ; Richardson et al., 2019 ; Spagnol et al., 2017 ; Stoll et al., 2015). Si le dispositif the vOICe n'a initialement pas été développé pour encoder la distance, il a été adapté pour le faire en utilisant une carte de profondeur en entrée encodée en nuances de gris (Hamilton-Fletcher et al., 2016b). C'est d'ailleurs ce qui est utilisé dans le dispositif MeloSee (Stoll et al., 2015). Pour améliorer la perception d'externalisation du son, certains de ces dispositifs ajoutent des indices de réverbération à la modulation du volume (Hamilton-Fletcher et al., 2022 ; Ribeiro et al., 2012).

On peut aussi mentionner les dispositifs qui utilisent des indices acoustiques temporels pour encoder la distance. Les dispositifs Colorphone (Bizoń-Angov et al., 2021) et SoundSight (Hamilton-Fletcher et al., 2022) utilisent un scanning temporel. Le dispositif See colOr module la durée du son en fonction de sa distance, alors que le dispositif GSSD de Paré et al. (2021) couple la modulation du volume avec celle de la vitesse de répétition du son (la vitesse augmentant à mesure que la distance diminue). Le dispositif GSSD module un troisième indice acoustique qui est la hauteur tonale, tout comme le dispositif LASS de Ton et al. (2018) qui utilise uniquement la modulation de la hauteur tonale (elle devient plus aigüe à mesure que la distance diminue).

3.1.4. Autres dimensions non-spatiales

Additionnellement aux dimensions spatiales que sont l'axe horizontal, l'axe vertical et la distance, certains DSSs encodent des dimensions visuelles telles que la luminosité (Cronly-Dillon et al., 1999 ; Hamilton-Fletcher et al., 2022 ; Hanneton et al., 2010 ; Meijer, 1992) et la couleur (Abboud et al., 2014 ; Bizoń-Angov et al., 2021 ; Deville et al., 2009 ; Hamilton-Fletcher et al., 2022 ; Hamilton-Fletcher et al., 2016a). Les indices acoustiques utilisés pour encoder la luminosité sont limitées au volume, alors que la couleur est encodée soit par la fréquence (Bizoń-Angov et al., 2021), soit par le timbre (Abboud et al., 2014 ; Deville et al., 2009 ; Hamilton-Fletcher et al., 2022 ; Hamilton-Fletcher et al., 2016a), ou le volume (Hamilton-Fletcher et al., 2022).

Si le dispositif SoundSight (Hamilton-Fletcher et al., 2022) propose d'encoder une dimension non-visuelle en soit, la température, en convertissant les images provenant d'une caméra thermique, la dimension visuelle résultante encodée est la couleur puisque les couleurs de l'image étaient associées à des températures.

3.2. Protocoles d'entraînement

3.2.1. Des entraînements actifs ou passifs

Maîtriser un DSS demande d'apprendre à interpréter les nouvelles informations auditives (détails dans la section II.1.3). Les protocoles d'entraînement utilisés jusqu'à présent varient par

leurs méthodes et leurs durées. On peut distinguer trois grandes familles de méthodes d'entraînement : les entraînements actifs, les entraînements passifs, et les explications verbales.

Dans les entraînements actifs, les participants sont exposés aux paysages sonores du DSS en explorant activement l'environnement autour d'eux en bougeant la caméra. Par exemple, les participants de l'étude de Pesnot Lerousseau et al. (2021) suivent un entraînement de 3 h réparties sur 2 sessions, comprenant une tâche durant laquelle ils apprennent à localiser un objet avec le dispositif the vOICe les yeux fermés. Dans Auvray et al. (2007), les participants suivent un protocole d'entraînement de 3 h. Après des explications verbales du schéma d'encodage du dispositif the vOICe, les participants suivent un entraînement pratique actif, consistant d'une part à explorer activement un objet statique avec le dispositif, puis à effectuer une tâche de suivi d'objet en mouvement. Les participants de l'étude de Renier et De Volder (2010) cumulent quant à eux environ 17 h d'entraînement avec le dispositif PSVA (dont 15 h d'entraînement et de pratique pour une tâche de reconnaissance de forme au cours de précédentes études). Des entraînements actifs de plus courtes durées peuvent être proposés comme dans Stoll et al. (2015) où l'entraînement ne dure que 8 min, comprenant des explications verbales et une exploration active de seulement 2 min durant laquelle les participants bougent librement avec le DSS. Commère et Rouat (2022) proposent une familiarisation active d'environ 10 min durant laquelle les participants peuvent explorer manuellement les objets à localiser et l'environnement autour d'eux en bougeant la tête. Si les entraînements mentionnés jusqu'à présent se déroulaient en laboratoire, Proulx et al. (2008) ont proposé un entraînement actif libre durant lequel les participants pouvaient utiliser librement le dispositif the vOICe pendant 21 jours, alors qu'ils étaient évalués sur 4 sessions. Les performances à la tâche de localisation par identification étaient meilleures en termes de rapidité et de précision que les participants n'ayant accès au dispositif que durant les phases de test. Les entraînements actifs ont l'avantage de permettre au participant d'acquérir les contingences sensorimotrices. La durée des entraînements actifs varie donc fortement, pouvant aller de 5 min (Pourghaemi et al., 2018) à des dizaines d'heures (Proulx et al., 2008). Dans les entraînements passifs, les participants sont exposés aux paysages sonores sans maîtriser la position et l'orientation de la caméra, ni la position des objets présents dans l'environnement. Avec des personnes voyantes, les stimuli visuels sont couramment présentés sur un écran en même temps que le paysage sonore associé (Ambard et al., 2015 ; Pesnot Lerousseau et al., 2021). Bien entendu, cette méthode d'entraînement a l'inconvénient de ne pas être adaptée aux personnes non-voyantes. Néanmoins, un entraînement reposant sur une exposition passive aux paysages sonores est parfois utilisé avec des personnes voyantes les yeux fermés (Abboud et al., 2014 ; Levy-Tzedek et al., 2012) ou des non-voyants (Abboud et al., 2014). Par exemple, le protocole de familiarisation de Levy-Tzedek et al. (2012) avec le dispositif EyeMusic consiste à diffuser successivement le paysage sonore associé

à une image alors que les participants ont les yeux bandés. Les participants ne pouvaient pas voir les images associées, mais le schéma d'encodage du dispositif leur était brièvement expliqué. Ces entraînements ont l'avantage d'être aisément implémentables, mais sont moins écologiques et peuvent être ennuyant, diminuant la motivation. Pour que le protocole de familiarisation soit utilisé auprès de personnes non-voyantes, il n'est bien évidemment pas envisageable de présenter les images correspondantes aux paysages sonores. Récemment, Buchs et al. (2021) ont proposé des entraînements passifs d'environ 1 h 15 avec le dispositif EyeMusic qui puissent être réalisés en autonomie. Ils ont comparé une méthode unisensorielle d'exposition aux paysages sonores (description verbale de l'image après la diffusion), avec des méthodes multisensorielles : une description écrite de l'image pendant la diffusion, ou bien l'affichage de l'image pendant ou après la diffusion du paysage sonore. Les quatre méthodes d'exposition permettaient d'améliorer les performances à une tâche de reconnaissance de forme, et un léger avantage était observé pour les méthodes multisensorielles. Leur étude suggère tout de même qu'une description verbale de l'image associée à un paysage sonore peut être utilisée comme méthode d'entraînement passif (d'un point de vue moteur) en autonomie chez des personnes non-voyantes.

À l'inverse, certaines études ne proposent pas d'entraînement en tant que tel mais se limitent à des explications verbales (Brown et al., 2011 ; Kim & Zatorre, 2008). Dans une première expérience, Kim et Zatorre (2008) ont comparé les performances de reconnaissance de formes avec the vOICe en fonction de l'entraînement. Un groupe de participants suivait un court entraînement passif, alors que l'autre groupe ne recevait que des explications verbales. Les deux groupes avaient des performances au-dessus du seuil de l'aléatoire et ne différaient pas l'un de l'autre, suggérant que le schéma d'encodage du dispositif the vOICe pouvait rapidement être interprété, même sans exposition. Dans une seconde expérience, (Kim & Zatorre, 2008) ont fait suivre un entraînement passif mais de façon intensive, sur 3 semaines, comprenant un total de 9 sessions d'entraînement d'environ 2 h, et une évaluation des performances par semaine. Parallèlement, un groupe contrôle réalisait uniquement l'évaluation hebdomadaire, sans explication verbale, ni feedback. Bien que les performances à la tâche de reconnaissance sur de nouveaux stimuli étaient inférieures à celles sur les stimuli familiers les performances s'amélioraient au fil des sessions uniquement pour le groupe ayant suivi l'entraînement intensif. Le groupe contrôle, bien que n'ayant reçu aucune explication verbale, avait des performances au-delà du seuil de l'aléatoire, suggérant une relative intuitivité du schéma d'encodage de the vOICe.

3.2.2. Le potentiel des entraînements en environnement virtuel

Lorsque les protocoles d'entraînement se déroulent dans un environnement réel, le contrôle de l'environnement et des différentes variables est plus difficile (Maidenbaum & Amedi,

2019). À l'inverse, si les protocoles d'entraînement sur écran permettent de standardiser les protocoles, de contrôler davantage de variables, et de proposer une adaptabilité progressive de la difficulté au fil de l'entraînement, ils présentent l'inconvénient d'être éloignés des situations réelles d'utilisation d'un dispositif de substitution. Suite à ce constat, le potentiel des environnements virtuels a été relevé dans plusieurs revues (Elli et al., 2014 ; Kristjánsson et al., 2016 ; Real et al., 2019), et démontré dans plusieurs études (Dascalu et al., 2017 ; Jicol et al., 2020 ; Maidenbaum & Amedi, 2019 ; Moldoveanu et al., 2017 ; Real & Araujo, 2021). Par exemple, dans Jicol et al. (2020), les participants se déplaçaient dans un environnement virtuel avec le dispositif the vOICe pour se familiariser avec le schéma d'encodage et établir une représentation mentale de l'espace physique dans lequel l'expérience allait se dérouler. Moldoveanu et al. (2017) ont proposé quant à eux un entraînement durant lequel les participants exploraient un environnement virtuel avec leur dispositif de substitution tout en restant assis mais en se déplaçant dans l'environnement à l'aide d'un joystick, ce que Maidenbaum et Amedi (2019) proposent de façon similaire avec le dispositif EyeMusic. Ainsi, les environnements virtuels permettent de conduire des protocoles d'entraînement (et d'évaluation) dans des contextes écologiques vis-à-vis d'une utilisation réelle d'un DSS mais sécurisés et contrôlés, tout en offrant une flexibilité graduelle dans le niveau de difficulté des tâches. Ces éléments permettent une standardisation des protocoles, favorisant la réplicabilité.

3.3. Capacités de localisation avec un dispositif de substitution

Les capacités de localisation avec un DSS ont été évaluées avec les dispositifs the vOICe (Auvray et al., 2007 ; Brown et al., 2011 ; Pesnot Lerousseau et al., 2021 ; Proulx et al., 2008), EyeMusic (Levy-Tzedek et al., 2012), Vibe (Hanneton et al., 2010), Synaestheatre (Richardson et al., 2019), See differently (Commere & Rouat, 2023), PSVA (Renier & De Volder, 2010), GSSD (Bazilinskyy et al., 2016) et ceux de Ambard et al. (2015), Guezou-Philippe et al. (2018), Mhaish et al. (2016) et Pourghaemi et al. (2018). Bien entendu, dans ces études, les participants voyants ont les yeux fermés. Ces études utilisent des tâches différentes, qui ne permettent pas toujours de comparer quantitativement les performances. Le **Tableau II-2** synthétise les études ayant évalué les capacités de localisation avec un DSS vision-vers-audition.

Tableau II-2. Synthèse d'études évaluant les capacités de localisation avec un DSS vision-vers-audition. Pour chaque étude, la tâche et la métrique utilisées sont spécifiées (le degré de liberté pour les tâches d'identification est précisé entre parenthèses) ainsi que les capacités de localisation en termes d'erreur de localisation lorsqu'elles ont été rapportées. Le type de stimuli visuel est précisé ainsi que le type d'entraînement et sa durée lorsqu'ils ont été rapportés, en différenciant explications verbales (Verbal), entraînement passif (sans contrôle de la caméra ou de l'environnement externe par le participant) et entraînement actif (le participant se familiarise avec le schéma d'encodage avec actions motrices).

Références **Nom du DSS**	**Tâche**	**Métrique**	**Erreur de localisation**	**Stimuli visuels**	**Entraînement**
Ambard et al. (2015)	Identification (1/9)	Taux de bonne réponse	.	Objet 2D sur un écran	Passif
Auvray et al. (2007) **The vOICe**	Pointage direct sur table	Erreur de localisation	Cartésienne : 7.8 ± 5.1 cm	Objet 3D réel	Verbal + actif (3 h)
Bazilinskyy et al. (2017) **GSSD**	Pointage direct sur écran (Direction) Echelle analogique [0, 100] (Distance)	Erreur de localisation	Azimut : entre 22.7 ± 7.5° et 24.5 ± 8.9° Distance : entre 11.8 ± 3.9 et 20.2 ± 7.6	Objet 2D sur un écran	.
Brown et al. (2011) **The vOICe**	Identification (1/9)	Précision (%)	.	Objets 3D réels	Non
Commère et al. (2020) **See differently**	Pointage direct sur table séquentiel (1-3 objets)	Erreur de localisation	Cartésienne : 2.1 ± 2.0 cm	Objets 3D réels	Actif (10 min)
Commère et al. (2022) **See differently**	Pointage direct sur table	Erreur de localisation	Cartésienne : 4.9 ± 3.5 cm	Objets 3D réels	Actif (5 à 10 min)
Commère et al. (2023)	Repositionner sur une table	Erreur de localisation	Distance : 12.3 ± 11.3 cm Azimut : entre ≈ 8° et 45°	Simulé	.

Tableau II.2. (suite)

Références **Nom du DSS**	**Tâche**	**Métrique**	**Erreur de localisation**	**Stimuli visuels**	**Entraînement**
Hanneton et al. (2010) **The Vibe**	Pointage direct sur table	Erreur de localisation	Cartésienne : 6.5 ± 5.0 cm Angulaire : 5.1 ± 4.5°	Non précisé	.
Levy-Tzedek et al. (2012) **EyeMusic**	Pointage direct sur écran	Erreur de localisation et taux de bonne réponse	Cartésienne : 0.4 ± 0.1 cm	Objet 2D sur écran	Verbal + passif (25 min)
Mhaish et al. (2016)	Identification (Direction : 1/4 ; Distance : 1/3)	Précision (%)	.	Objet 3D réel	Verbal + Actif (5 min)
Pesnot Lerousseau et al. (2021) **The vOICe**	Pointage direct sur table	.	.	Objet 2D réel	Actif + Passif (3 h)
Pourghaemi et al. (2018)	Pointage direct sur table	Score de précision {Précis ; Faible ; Echec}	.	Objet 3D réel	Verbal + Actif (45 min)
Proulx et al. (2008) **The vOICe**	Identification (1/18)	Taux de bonne réponse	.	Objet 2D réel	Actif (de 0 à 21 jours)
	Identification (6/164)				
	Attraper sur une table	Score {1 ; 2 ; 3}		Object 3D réel	
Renier et al. (2010) **PSVA**	Attraper sur une table	Erreur de localisation	Distance : entre 11 et 19 cm Latérale : entre 5 et 5.5 cm	Objet 3D réel	Actif (2 h+15 h)
Richardson et al. (2019) **Synaestheatre**	Discrimination (1/2)	Score de discrimination (Élévation)	Cartésienne : 30.1 ± 18.9 cm Elévation : 14.3°	Objet 3D réel	.
		Score de discrimination (Distance)	Cartésienne : 8.3 ± 7.3 cm		

Parmi les études, on retrouve l'utilisation de tâches de pointage (Auvray et al., 2007 ; Bazilinskyy et al., 2016 ; Commère et al., 2020 ; Commère & Rouat, 2022, 2023 ; Hanneton et al., 2010 ; Levy-Tzedek et al., 2012 ; Pesnot Lerousseau et al., 2021 ; Pourghaemi et al., 2018 ; Proulx et al., 2008 ; Renier & De Volder, 2010), d'identification de position (Ambard et al., 2015 ; Brown et al., 2011 ; Mhaish et al., 2016 ; Proulx et al., 2008), et de discrimination (Commère & Rouat, 2023 ; Richardson et al., 2019).

Dans les tâches de pointage, le participant doit pointer (parfois attraper ou repositionner) soit un objet 3-dimensionnel sur une table, soit un stimulus 2-dimensionnel sur un écran. Cette méthode présente l'avantage de pouvoir mesurer quantitativement une performance de localisation absolue (Auvray et al., 2007 ; Bazilinskyy et al., 2016 ; Commère et al., 2020 ; Commère & Rouat, 2022, 2023 ; Hanneton et al., 2010 ; Levy-Tzedek et al., 2012 ; Renier & De Volder, 2010), bien qu'elle ne soit pas toujours mesurée. Elle est aussi écologique puisqu'il s'agit d'une tâche qu'une personne non-voyante peut vouloir effectuer avec l'aide d'un DSS. Néanmoins, la tâche présente la contrainte de devoir présenter les objets à des distances proches, voire atteignables. De plus, pour être facilement réplicable, cela demande d'utiliser des stimuli visuels et des environnements expérimentaux comparables.

En termes de performances, lorsque les tâches consistaient à pointer des objets réels 3-dimensionnel, la distance moyenne entre la position de l'objet et celle de la réponse était comprise entre 2 et 19 cm en fonction du DSS utilisé. Avec le dispositif the vOICe, une erreur moyenne de pointage de l'ordre de 8 cm a été observée (Auvray et al., 2007), ce qui est comparable à celle d'environ 7 cm avec le dispositif Vibe (Hanneton et al., 2010) et entre 5 et 11 cm avec PSVA (Renier & De Volder, 2010). Avec See Differently, Commère et al. (2020) et Commère et Rouat (2022) rapportent des erreurs comprises entre 2 et 5 cm. En utilisant une tâche de pointage sur un écran avec le dispositif EyeMusic, Levy-Tzedek et al. (2012) mesurent un erreur moyenne inférieure à 1 cm. Les études mesurant l'erreur angulaire en azimut dans la tâche de pointage observent des erreurs angulaires entre 8 et 45° en fonction du schéma d'encodage utilisé (Commère & Rouat, 2023), autour de 5° avec Vibe (Hanneton et al., 2010) et 25° avec GSSD (Bazilinskyy et al., 2016).

Les tâches de discrimination spatiale utilisées consistent à présenter deux stimuli visuels (perçus avec le paysage sonore du DSS) que le participant doit comparer pour déterminer quel stimulus est, par exemple, le plus distant (Commère & Rouat, 2023 ; Richardson et al., 2019). Ce type de tâche présente l'avantage de mesurer quantitativement une performance de localisation avec un DSS, mais cette fois ci la perception spatiale est relative. Un seuil de discrimination est alors déterminé, en utilisant par exemple une méthode staircase qui consiste à réduire et à augmenter à plusieurs reprises l'écart entre les deux stimuli. Les tâches de discrimination sont moins

répandues, mais tout de même existantes avec l'étude de Richardson et al. (2019) qui étudient les seuils de discrimination de distance et d'élévation avec le dispositif Synaestheatre. Ils mesurent avec cette méthode un seuil de discrimination d'environ 8 cm pour la distance, et de 15° pour l'élévation. En utilisant l'équivalent d'objets virtuels simulés avec un DSS, Commère et Rouat (2023) mesurent des seuils de discrimination allant de 1 à 9 cm en fonction du schéma d'encodage et de la distance de référence.

Enfin, dans les tâches d'identification de position (Ambard et al., 2015 ; Brown et al., 2011 ; Mhaish et al., 2016 ; Proulx et al., 2008), le participant doit identifier la position d'un stimulus avec le DSS, parmi un effectif discret de positions au choix. La plupart du temps, ces tâches permettent d'évaluer les performances de localisation avec un taux de bonnes réponses, qui doit donc être interprété au regard du nombre de positions possible.

3.4. Capacités à utiliser un dispositif de substitution dans une scène complexe

En contexte d'utilisation écologique d'un DSS, l'environnement visuel est bien souvent riche, et le paysage sonore du DSS est délivré parmi des informations auditives réelles qui nécessitent elles aussi d'être traitées. Il est alors nécessaire d'être capable d'interpréter les informations provenant du DSS ainsi que les informations de l'environnement. Si cet élément fondamental est couramment mis en avant lors d'études sur les DSSs, il n'a que peu été testé expérimentalement. Il semblerait qu'un environnement extérieur complexe et bruyant rende plus difficile l'interprétation du paysage sonore généré par le DSS (Elli et al., 2014).

Les capacités à effectuer une tâche d'identification d'objets avec le dispositif EyeMusic (Abboud et al., 2014) alors que des informations auditives réelles mais non-pertinentes étaient diffusées, ont été évaluées par Buchs et al. (2019). Dans leur étude, les participants devaient reconnaître des objets avec le dispositif, alors que des stimuli auditifs environnementaux (e.g., des gens qui discutent) étaient diffusés dans la pièce simultanément. Pour étudier l'effet de ces stimuli environnementaux, ils ont comparé les performances à la tâche avec EyeMusic avec et sans fond sonore. Leurs résultats ne montrent pas de diminution des performances à la tâche avec la présence du fond sonore. Les participants arrivaient donc bien à interpréter les informations auditives pertinentes à la tâche (i.e., celle du DSS) et ignorer celles qui ne l'étaient pas (i.e., bruits environnementaux). Notons que les participants étaient des personnes non-voyantes congénitales ayant une expérience préalable avec le dispositif EyeMusic. À noter aussi que les stimuli auditifs réels diffusés n'étaient pas pertinents pour la tâche en cours.

Or, dans un contexte d'utilisation réelle (e.g., un déplacement urbain), il est essentiel de percevoir les bruits environnants, ce qui s'apparente à une double tâche, impliquant une charge cognitive supérieure. Dans une tâche de navigation, Stoll et al. (2015) ont proposé aux participants de se déplacer en utilisant le DSS MeloSee, tout en effectuant une tâche distractrice de discrimination tactile. Leurs résultats démontrent que la charge cognitive induite par la double tâche n'augmentait pas le nombre de contact avec les obstacles, et augmentait peu le temps de parcours, et seulement lorsque le trajet était le plus compliqué. Dans le contexte d'une tâche de navigation avec un guidage sonore, Klatzy et al. (2006) ont cherché à reproduire une situation de double tâche pour mesurer la charge cognitive induite par deux types de guidage sonore : un son virtuel spatialisé en azimut ou des indications verbales spatiales (i.e. « gauche » ou « droite »). Notons qu'il ne s'agit pas de substitution sensorielle en tant que telle. Dans une tâche de navigation les yeux fermés avec le système de guidage, les participants devaient simultanément effectuer une tâche de N-back vibrotactile. Leurs résultats montrent qu'en situation de charge cognitive (i.e., de double tâche), les performances dans la tâche de navigation diminuaient uniquement avec le guidage utilisant les indications verbales, et non avec celui utilisant le son spatialisé, suggérant un avantage de l'utilisation de sons spatialisés en situation de charge cognitive. Ainsi, il apparaît essentiel de limiter la charge cognitive et perceptive induite par l'utilisation du DSS, notamment dans des environnements peu sécurisés comme lors d'un déplacement pédestre en milieu urbain.

De plus, en contexte d'utilisation réelle d'un DSS, pour qu'une représentation mentale de la scène visuelle puisse être construite, il faut être en capacité de ségréguer les informations contenues dans le paysage sonore du DSS. Avec le dispositif the vOICe, qui module la hauteur tonale pour encoder l'axe vertical de l'image, Brown et al. (2015) ont montré que les capacités à percevoir deux lignes horizontales distinctes avec le dispositif étaient influencées par les positions relatives des deux lignes. En effet, les capacités étaient altérées lorsque les deux lignes étaient associées à des fréquences consonantes, menant à une fusion des deux tonalités dans le paysage sonore, donc des deux lignes dans l'image. L'étude de Brown et al. (2015) mettait donc en évidence une limite de perception avec le dispositif the vOICe inhérente aux principes de l'analyse de scènes auditives. Dans ce contexte, une étude plus récente met en évidence que les principes de l'analyse de scènes auditives peuvent faciliter les capacités à ségréguer le paysage sonore du dispositif the vOICe pour établir une représentation mentale des différents éléments présents dans une image (Hamilton-Fletcher et al., 2021). Dans leur étude, les auteurs montraient que moduler également le timbre ou l'intensité en plus de la hauteur tonale pour l'axe vertical pouvait parfois améliorer les capacités de ségrégation du paysage sonore. Ces deux études démontrent que les capacités d'utilisation d'un DSS sont influencées à la fois par le schéma d'encodage utilisé et par la configuration de la scène (ou image) perçue à travers le paysage sonore.

3.5. Synthèse

- Les DSSs existants utilisent différentes méthodes de conversion, en combinant la reproduction d'indices acoustiques spatiaux mais aussi en utilisant d'autres dimensions du son.

- La reproduction d'indices spatiaux acoustiques se place dans une approche d'équivalence fonctionnelle auditive et fait donc face aux mêmes limites que celles des capacités de perception spatiale auditive.

- À l'inverse, la modulation d'autres dimensions du son paraît davantage arbitraire, bien qu'elle repose parfois sur des correspondances cross-modales.

- Pour apprendre à interpréter les indices acoustiques provenant d'un DSS, des protocoles d'entraînement actifs et passifs peuvent être mis en place. Les environnements virtuels peuvent permettre de standardiser les protocoles d'évaluation et d'entraînement en contrôlant davantage les variables, et offrent une flexibilité graduelle permettant d'adapter les niveaux de difficulté des tâches en modulant l'environnement virtuel.

- La localisation d'objets sans la vision est bel et bien rendue possible avec divers DSSs dans des environnements minimalistes, mais la diversité des tâches limite la comparaison des capacités entre les dispositifs.

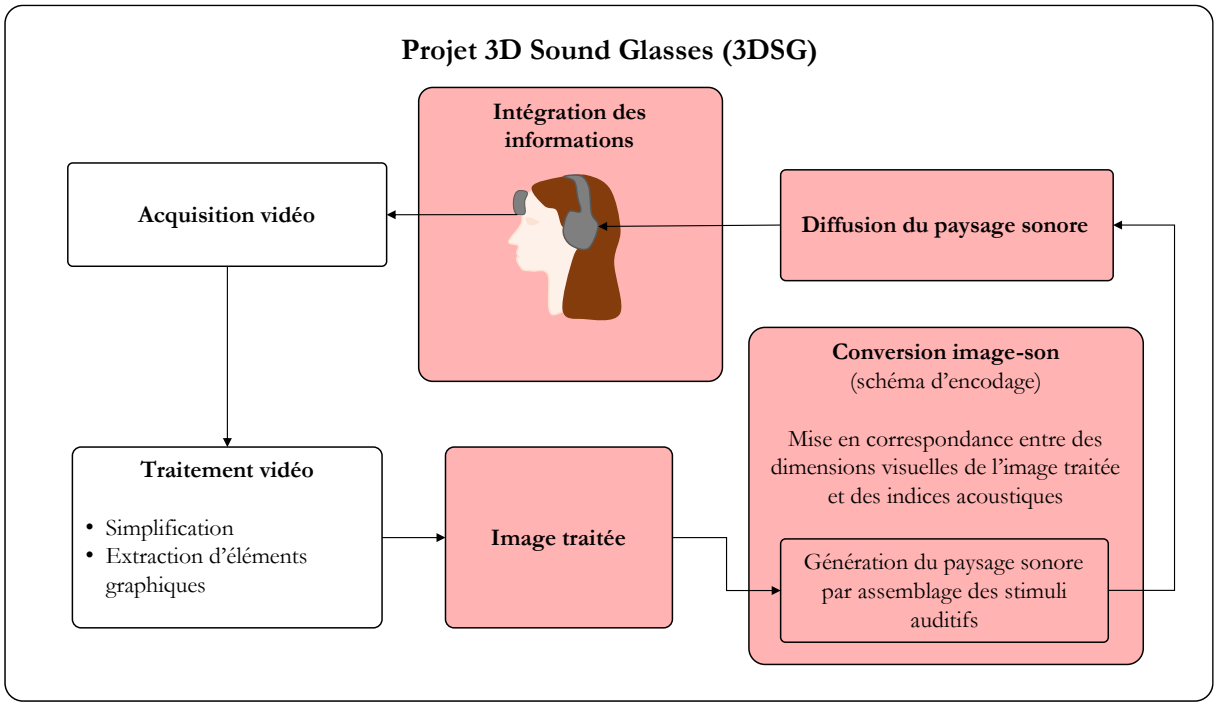
- Les capacités d'utilisation d'un DSS dans un environnement plus riche ont été peu étudiées, et sont préservées dans une certaine mesure.

Dans le contexte du développement d'un DSS vision-vers-audition pour l'aide à la locomotion et à la localisation d'obstacles, le manque de comparaison des capacités de localisation entre les DSSs ne permet pas de déterminer si les capacités de perception spatiale avec un DSS sont facilitées par l'utilisation d'indices acoustiques spatiaux ou par la modulation d'autres indices acoustiques. Bien que les capacités de perception avec un DSS dans un environnement plus riche puissent être préservées, les indices acoustiques utilisés par le DSS devraient influencer ces capacités, et les limites perceptives spatiales n'ont que très peu été étudiées à ce jour.

III. Problématique

Les dispositifs de substitution (DSSs) vision-vers-audition représentent un enjeu sociétal d'inclusion important pour aider les personnes non-voyantes à gagner en autonomie. Leur objectif est de faciliter l'intégration d'informations de la scène environnante par le biais de la modalité auditive. La présente thèse s'intègre au projet 3D Sound Glasses (3DSG) qui vise à développer un dispositif de substitution sensorielle vision-vers-audition pour l'aide à la locomotion et la localisation d'obstacles. Pour concevoir un tel système, il est primordial de prendre en compte les capacités humaines perceptives auditives et spatiales pour déterminer les indices acoustiques à utiliser dans le schéma d'encodage, afin de faciliter l'apprentissage et limiter la surcharge cognitive et perceptive. Dans le processus de développement du DSS du projet 3DSG, les travaux de la présente thèse se concentrent sur les dernières étapes du schéma d'encodage de l'information que sont la conversion de l'image traitée en sons, la diffusion du paysage sonore et l'intégration de ces informations auditives (**Figure III-1**).

Figure III-1. Aspects sur lesquels la présente thèse se focalise dans le processus de développement du DSS dans le projet 3DSG (en rouge).



Les travaux de cette thèse ont pour objectif de comparer et déterminer des indices acoustiques pouvant être utilisés dans le DSS en prenant en compte les capacités perceptives humaines ainsi que la pertinence de ces indices acoustiques pour la fonctionnalité recherchée qui est dans notre cas l'aide à la locomotion et à la localisation d'obstacles. La question au cœur de ce travail de thèse concerne le choix des dimensions du son à moduler pour transmettre des informations spatiales avec un DSS ? Les capacités de localisation avec le dispositif peuvent-elles être facilitées par l'utilisation d'indices acoustiques sur lesquels le système auditif se repose pour

localiser une source sonore ? Au contraire, l'utilisation d'indices acoustiques impliqués dans des effets d'interaction audio-visuelle peut-elle améliorer les capacités de perception spatiale avec le dispositif ? Finalement, dans quelle mesure les indices acoustiques rendent possible la localisation d'obstacles lorsque la scène est complexe ?

Pour aborder ces questions, le travail de thèse présenté dans ce document s'articule autour de deux axes qui sont schématisés dans la **Figure III-2**. Le premier axe vise à déterminer des indices acoustiques pouvant être utilisés dans le dispositif de substitution sensorielle pour transmettre des informations spatiales 2-dimensionnelles (**Étude 1**) et 3-dimensionnelles (**Étude 2**) dans un environnement virtuel minimaliste. Le deuxième axe vise à étudier dans quelle mesure les capacités de localisation peuvent être altérées dans un environnement plus complexe comprenant des objets distracteurs (**Étude 3**).

Figure III-2. Plan de la thèse. Le premier axe (en vert) est abordé par l'Étude 1 et l'Étude 2, et le deuxième axe (en orange) est abordé par l'Étude 3.

Pour évaluer quantitativement et précisément les capacités de perception spatiale avec le dispositif dans un environnement contrôlé et sécurisé, facilitant la réplicabilité, des tâches expérimentales ont dû être développées dans un environnement virtuel. Puisque la durée des entraînements et la non-intuitivité des indices acoustiques sont des freins à l'adoption des dispositifs, les capacités d'utilisation du dispositif sont étudiées aux premiers stades de son utilisation. Les dispositifs devant pouvoir être évalués chez la population cible non-voyante, les protocoles de familiarisation et les tâches de localisation doivent être développées pour être facilement adaptables avec des personnes non-voyantes.

La première étude a pour objectif d'évaluer les capacités de localisation de cibles virtuelles avec le DSS dans les dimensions de l'azimut et de l'élévation en comparant trois schémas d'encodage. Deux types d'indices acoustiques pour l'élévation sont comparés : la spatialisation avec HRTFs non-individualisées en azimut et en élévation appliquée soit à un bruit blanc, soit à des tonalités (pures ou complexes) permettant de coupler la spatialisation avec une modulation de la hauteur tonale pour l'élévation. Du fait des limites perceptives spatiales pour l'élévation avec des HRTFs non-individualisées, nous émettons l'hypothèse que la modulation de la hauteur tonale facilitera les capacités de localisation en comparaison à la spatialisation d'un bruit blanc, notamment grâce à un effet facilitateur de la correspondance cross-modale audio-visuelle entre la hauteur tonale et l'élévation. Nous émettons également l'hypothèse que les capacités de localisation de l'élévation avec le schéma d'encodage intégrant la modulation de la hauteur tonale seront facilitées lorsque les tonalités sont complexes.

La seconde étude a pour objectif d'évaluer les capacités de localisation de la distance de cibles virtuelles avec le DSS en comparant deux schémas d'encodage pour la distance : l'un utilisant seulement la modulation de l'intensité en se rapprochant de l'atténuation de l'intensité observée lors de la propagation de sources sonores réelles, et l'autre intégrant aussi une modification de l'enveloppe, changeant le timbre. Cette étude a également pour objectif de proposer un nouveau protocole pour évaluer les capacités de perception de la distance avec un DSS. Nous faisons l'hypothèse qu'un biais de compression de la distance similaire à celui mesuré dans la perception spatiale auditive sera observé. Également, nous faisons l'hypothèse que la modification de l'enveloppe sonore, connue comme pouvant interagir avec la perception spatiale, peut moduler les capacités d'estimation de la distance.

Après avoir déterminé les indices acoustiques à utiliser dans le schéma d'encodage du DSS pour les 3 dimensions (azimut, élévation et distance), la troisième étude a pour objectif d'évaluer les capacités de localisation dans un environnement plus complexe se rapprochant des conditions réelles d'utilisation. L'objectif de cette troisième étude est d'évaluer les capacités de localisation

d'une cible virtuelle avec le dispositif alors qu'il est présenté parmi d'autres objets virtuels à ignorer. En comparant les capacités de localisation avec et sans distracteurs, et en fonction de leur arrangement spatial, l'objectif est d'évaluer les limites des indices acoustiques utilisés dans le dispositif lorsque les objets sont à localiser dans une scène complexe. Nous faisons l'hypothèse que la présence de distracteurs entraînera une diminution des capacités de localisation, mais que l'utilisation de la hauteur tonale comme indice acoustique facilitera la ségrégation du paysage sonore pour localiser la cible virtuelle parmi les distracteurs.

IV. Partie expérimentale

1. Étude 1

La correspondances cross-modale améliore les performances de localisation de l'élévation avec un dispositif de substitution sensorielle vision-vers-audition

Cross-modal correspondence enhances elevation localization in visual-to-auditory sensory substitution

Camille Bordeau, Florian Scalvini, Cyrille Migniot, Julien Dubois and Maxime Ambard

Publié dans *Frontiers in Psychology (2023)*

https://doi.org/10.3389/fpsyg.2023.1079998

Mis en format article en **Annexe A**

1.1. Résumé

Les dispositifs de substitution sensorielle (*SSDs* pour *Sensory Substitution Devices* en anglais) vision-vers-audition sont des dispositifs d'assistance pour les personnes non-voyantes qui convertissent les images visuelles en images auditives (ou paysages sonores) en associant des caractéristiques visuelles à des indices acoustiques. Pour transmettre des informations spatiales à l'aide de sons, plusieurs SSDs utilisent un espace acoustique virtuel (VAS *pour Virtual acoustic space* en anglais) à l'aide des fonctions de transfert HRTFs (pour *Head Related Transfer Functions* en anglais) pour synthétiser les indices acoustiques naturels utilisés pour la localisation de sources sonores. Cependant, la perception de l'élévation est connue pour être inexacte avec lorsque des HRTFs non-individualisées sont utilisées pour la spatialisation, puisqu'elle est basée sur des modifications spectrales qui sont spécifiques à chaque individu. Une autre méthode utilisée pour transmettre des informations sur l'élévation est basée sur la correspondance audiovisuelle entre la hauteur tonale et l'élévation visuelle. Le principal inconvénient de cette deuxième méthode est la limitation de la capacité à percevoir l'élévation par le biais des HRTFs en raison de la bande spectrale étroite des sons.

Dans cette étude, nous avons comparé la capacité à localiser des objets à l'aide d'un SSD où l'élévation est transmise soit en utilisant une méthode basée uniquement sur la spatialisation (encodage *Noise*), soit en utilisant des méthodes reposant sur la modulation de la fréquence en comparant différentes complexités spectrales (encodages *Monotonic* et *Harmonic*). Trente-huit participants aux yeux bandés ont dû localiser une cible virtuelle en utilisant des paysages sonores provenant du SSD, avant et après avoir été familiarisés avec les méthodes d'encodage.

Les performances de localisation de l'élévation de la cible virtuelle étaient supérieures avec les encodages reposant sur la modulation de la fréquence qu'avec l'encodage basé uniquement sur la spatialisation. Seules de légères différences dans les performances de localisation de l'azimut ont été constatées entre les encodages. Cette étude suggère l'intuitivité d'un encodage basé sur la modulation de la fréquence avec un effet de facilitation de la correspondance cross-modale audiovisuelle lorsqu'une méthode de spatialisation sonore non-individualisée est utilisée.

1.2. Article

Cross-modal correspondence enhances elevation localization in visual-to-auditory sensory substitution

Camille Bordeau[1], Florian Scalvini[2], Cyrille Migniot[2], Julien Dubois[2] and Maxime Ambard[1]

[1] LEAD-CNRS UMR5022, Université de Bourgogne, Dijon, France
[2] ImViA EA 7535, Université de Bourgogne, Dijon, France

Abstract

Introduction: Visual-to-auditory sensory substitution devices are assistive devices for the blind that convert visual images into auditory images (or soundscapes) by mapping visual features with acoustic cues. To convey spatial information with sounds, several sensory substitution devices use a Virtual Acoustic Space (VAS) using Head Related Transfer Functions (HRTFs) to synthesize natural acoustic cues used for sound localization. However, the perception of the elevation is known to be inaccurate with generic spatialization since it is based on notches in the audio spectrum that are specific to each individual. Another method used to convey elevation information is based on the audiovisual cross-modal correspondence between pitch and visual elevation. The main drawback of this second method is caused by the limitation of the ability to perceive elevation through HRTFs due to the spectral narrowband of the sounds.

Method: In this study we compared the early ability to localize objects with a visual-to-auditory sensory substitution device where elevation is either conveyed using a spatialization-based only method (Noise encoding) or using pitch-based methods with different spectral complexities (Monotonic and Harmonic encodings). Thirty-eight blindfolded participants had to localize a virtual target using soundscapes before and after having been familiarized with the visual-to-auditory encodings.

Results: Participants were more accurate to localize elevation with pitch-based encodings than with the spatialization-based only method. Only slight differences in azimuth localization performance were found between the encodings.

Discussion: This study suggests the intuitiveness of a pitch-based encoding with a facilitation effect of the cross-modal correspondence when a non-individualized sound spatialization is used.

Keywords: Virtual Acoustic Space, spatial hearing, sound spatialization, image-to-sound conversion, cross-modal correspondence, assistive technology, visual impairment, sound source localization

1. Introduction

Visual-to-auditory Sensory substitution devices (SSDs) are assistive tools for blind people. They convert visual information into auditory information in order to convey spatial information about the surrounding environment when vision is impaired. The visual-to-auditory conversion relies on the mapping of selected visual features with specific auditory cues. Visual information is usually acquired using a camera capturing the visual scene in front of the person. Then the scene converted into auditory information is transmitted to the user through soundscapes (or auditory images) delivered with headphones.

Various visual-to-auditory encodings are used by the existing visual-to-auditory SSDs to convey spatial information. Some of them use encoding schemes based on a Virtual Acoustic Space (VAS). A VAS consists in the simulation of a binaural acoustic signature of a virtual sound source located in a 3D space. In the context of visual-to-auditory SSDs, this is mainly used to simulate sound sources at the location of the obstacles. This simulation is achieved by spatializing the sound through the incorporation of spatial auditory cues in the original monophonic sound. Then a synthesized stereophonic signal simulating the distortions occurring while receiving the audio signal by the two ears is obtained. Among the SSDs used in localization experiments, the Synaestheatre (Hamilton-Fletcher et al., 2016a; Richardson et al., 2019), the Vibe (Hanneton et al., 2010) and the one presented by Mhaish et al. (2016) spatialize azimuth (lateral position) and elevation (vertical position). Other SSDs only spatialize the azimuth: the See differently device (Rouat et al., 2014), the one studied in Ambard et al. (2015), and the recent one presented in Scalvini et al. (2022).

The generation of a VAS is based on the reproduction of binaural acoustic cues related to the relative sound source location such as timing, intensity and spectral features (for an in-depth explanation of the auditory localization mechanisms see Blauert, 1996). Those features arise from audio signal distortions mainly caused by the reflection and absorption of the head, pinna and torso and are partly reproducible using Head-Related Transfer Functions (HRTFs). HRTFs are transfer functions characterizing these signal distortions as a function of the position of the sound source relatively to the two ears. They are usually obtained by conducting multiple binaural recordings with a sound source carefully placed in various positions while repeatedly producing the same sound.

Due to the technical difficulty in acquiring these recordings in good conditions, non-individualized HRTFs acquired in controlled conditions with another listener or a manikin are frequently used. However, these HRTFs failed to simulate the variability of individual-specific spectrum distortions that are related to individual morphologies (head, torso and pinna). Consequently, the localization of simulated sound sources using non-individualized HRTFs is often

inaccurate with front-back and up-down confusions that are less resolvable (Wenzel et al., 1993), and a less perceptible externalization (Best et al., 2020). Nonetheless, due to the robustness of the binaural cues, azimuth localization accuracy is well preserved compared to the perception of elevation since azimuth perception relies less on the individual-specific spectrum distortions (Makous and Middlebrooks, 1990; Wenzel et al., 1993; Middlebrooks, 1999). Therefore, visual-to-auditory encodings only based on the creation of a VAS have the advantage to rely on acoustic cues that mimic natural acoustic features for sound source localization, nevertheless in practice the elevation perception can be impaired.

To compensate for this difficulty some visual-to-auditory SSDs use additional acoustic cues to convey spatial information. For instance, pitch modulation is often used to convey elevation location (Meijer, 1992; Abboud et al., 2014; Ambard et al., 2015). This mapping between elevation location and auditory pitch is based on the audiovisual cross-modal correspondence between pitch and elevation (see Spence, 2011 for a review on audiovisual cross-modal correspondences). Humans show a tendency to associate high pitch with high spatial locations and low pitch with low spatial locations. For example, they tend to exhibit faster response times in an audio-visual Go/No-Go task when the visual and auditory stimuli are congruent, i.e., higher pitch with higher visual location, and lower pitch with lower visual location (Miller, 1991). They also tend to discriminate more accurately and quickly the location of a visual stimulus (high vs. low location) when the pitch of a presented sound is congruent with the visual elevation (Evans and Treisman, 2011). Also, humans tend to respond to high pitch sounds with a high-located response button instead of a lower-located response button (Rusconi et al., 2006). The pitch-based encoding used in the vOICe SSD (Meijer, 1992) has been suggested somewhat intuitive in a recognition task (Stiles and Shimojo, 2015). Nevertheless, the main drawback of a pitch-based encoding is caused by the limitation of the abilities to perceive elevation through HRTFs due to the audio spectral narrowband (Algazi et al., 2001b). Although some acoustic cues for elevation perception are present in low frequencies below 3,500 Hz (Gardner, 1973; Asano et al., 1990), localization abilities are higher when the spectral content contains high frequencies above 4,000 Hz (Hebrank and Wright, 1974; Middlebrooks and Green, 1990). Since the ability to perceive the elevation through HRTFs is higher with broadband sounds containing high frequencies, the spectral content of the sound used in the visual-to-auditory encoding might modulate the perception of elevation through HRTFs. No study has directly compared encodings only based on HRTFs with encodings adding a pitch modulation and it remains unclear if the simulation of natural acoustic cues is less efficient for object localization than a more artificial sonification method using the cross-modal correspondence between pitch and elevation.

Many studies investigating static object localization abilities have already been conducted with blindfolded sighted persons using visual-to-auditory SSDs. Various types of tasks have already been used, for example discrimination tasks with forced choice (Proulx et al., 2008; Levy-Tzedek et al., 2012; Ambard et al., 2015; Mhaish et al., 2016; Richardson et al., 2019), grasping tasks (Proulx et al., 2008), index or tool pointing tasks (Auvray et al., 2007; Hanneton et al., 2010; Brown et al., 2011; Pourghaemi et al., 2018; Commère et al., 2020), or head-pointing tasks (Scalvini et al., 2022). Those studies showed the high potential of SSDs to localize an object and interact with it. However, long trainings were often conducted before the localization tasks to learn the visual-to-auditory encoding schemes: from 5 min in Pourghaemi et al. (2018) to 3 h in Auvray et al. (2007). On the contrary, in the study of Scalvini et al. (2022) the experimenter only explained verbally the encoding schemes to the participants.

Virtual environments are more and more used to investigate the abilities to perceive the environment with a visual-to-auditory SSD (Maidenbaum et al., 2014; Kristjánsson et al., 2016) since they allow a complete control of the experimental environment (e.g., number of objects, object locations...) (Maidenbaum and Amedi, 2019) and a more accurate assessment of localization abilities with precise pointing methods. They have been used in standardization tests to compare the abilities to interpret information provided by SSDs in navigation or localization tasks (Caraiman et al., 2017; Richardson et al., 2019; Jicol et al., 2020; Real and Araujo, 2021).

The current study aimed at investigating the intuitiveness of different types of visual-to-auditory encodings for the elevation in the context of object localization with a SSD. Therefore, we conducted a localization task in a virtual environment with blindfolded participants testing a spatialization-based encoding and a pitch-based encoding. This study also aimed at assessing whether a higher spectral complexity of the sound used in a pitch-based encoding could improve the localization performance. Therefore, 2 types of pitch-based encodings were investigated: one monotonic and one harmonic with 3 octaves. We measured the localization performance for the azimuth and for the elevation. For each of these measures, we studied the effect of the visual-to-auditory encoding before and after an audio-motor familiarization of short duration.

Since the audio spatialization method was not based on individualized HRTFs, and since the pitch-based encodings were not explained to the participants, localization performance for the elevation was expected to be impaired. However, a facilitation effect of the pitch-based encodings for the elevation localization accuracy was hypothesized. Among the two pitch-based encodings, a higher elevation localization accuracy was predicted with the harmonic encoding since the sound has a higher spectral complexity. Also, the intuitiveness of the azimuth perception for all the

encodings was hypothesized since it is based on less individual-specific acoustic spatial cues than elevation perception.

2. Method

2.1 Participants

Thirty eight participants were divided into two groups: the Monotonic group (19, age: M = 25.5, SD = 3.04, 6 female, 19 right-handed) and the Harmonic group (19, age: M = 24.4, SD = 3.27, 10 female, 18 right-handed). No participant reported impairments of hearing or any history of psychiatric illness or neurological disorder. The experimental protocol was approved by the local ethical committee Comité d'Ethique pour la Recherche de Université Bourgogne Franche-Comté (CERUBFC-2021-12-21-050) and followed the ethical guidelines of the Declaration of Helsinki. Written informed consent was obtained from all the participants before the experiment. No monetary compensation was given to the participants.

2.2 Visual-to-auditory conversion in the virtual environment

The visual-to-auditory SSD used took place in a virtual environment created in UNITY3D and including the target to localize, a virtual camera, and a tracked pointing tool. Four HTC VIVE base stations were used to track the participants' head and the pointing tool on which HTC VIVE Trackers 2.0 were attached. Participants did not carry a headset and therefore could not explore visually the virtual environment. The pointing task can be separated in several steps that are explained in detail below: the virtual target placement, the video acquisition from a virtual camera, the video processing, the visual-to-auditory conversion and the participants' response collection using the pointing tool.

2.2.1 Virtual target

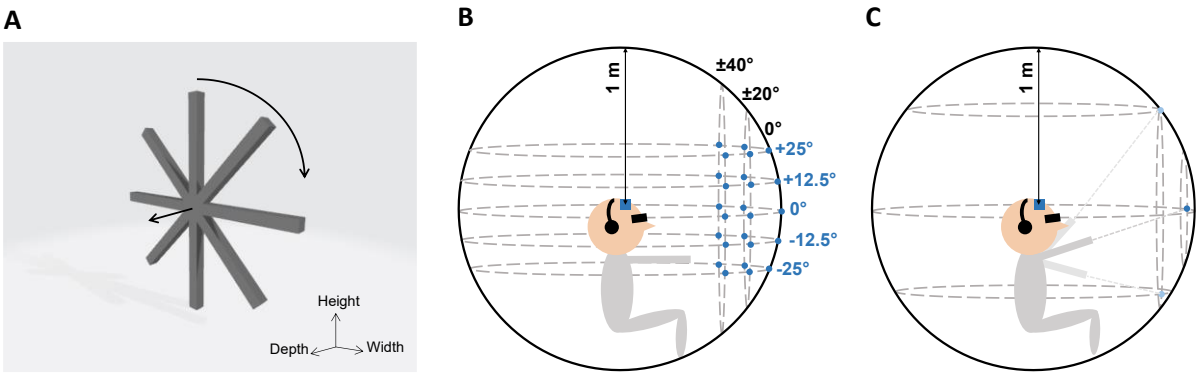
The virtual target that participants had to localize was a 3D propeller shape of 25 cm in diameter composed of 4 bars with a length of 25 cm and a rectangular section of 5 × 5 cm that was self-rotating at a speed of 10° per video frame (see Figure 1A). The use of an angular shaped target that is self-rotating generated a modification of successive video frames without changing the center position of the target. The orientation of the target was managed in order to continuously face the virtual camera while being displayed. Since participants could not see the virtual target, it was only perceivable through the soundscapes.

2.2.2 Video acquisition

The virtual camera position was set at the beginning of each trial using the position of the head tracker attached on the participants' forehead. Images were acquired with a virtual camera

with a field of view of 90 × 74° (Horizontal × Vertical) and a frame rate of 60 Hz. The resulting image was using a grayscale encoding (0–255 gray levels) of a depth map (0.2 m = 0, 5.0 m = 255) of the virtual scene although in this experiment we did not manipulate the depth parameter.

Figure 1. (**A**) The virtual target was a self-rotating 3D propeller shape. The straight arrow shows the forward axis facing the virtual camera. The circular arrow shows the self-rotation direction. (**B**) In each localization test, 25 target positions (blue circles) were tested, including 5 elevation positions (horizontal dotted ellipses) tested at 5 azimuth positions (vertical dotted ellipses). (**C**) In the familiarization session, participants placed the virtual target during 60 seconds on the grid by moving the pointing tool. Participants' head were tracked during the localization tests and familiarization sessions (blue square).



2.2.3 Video processing

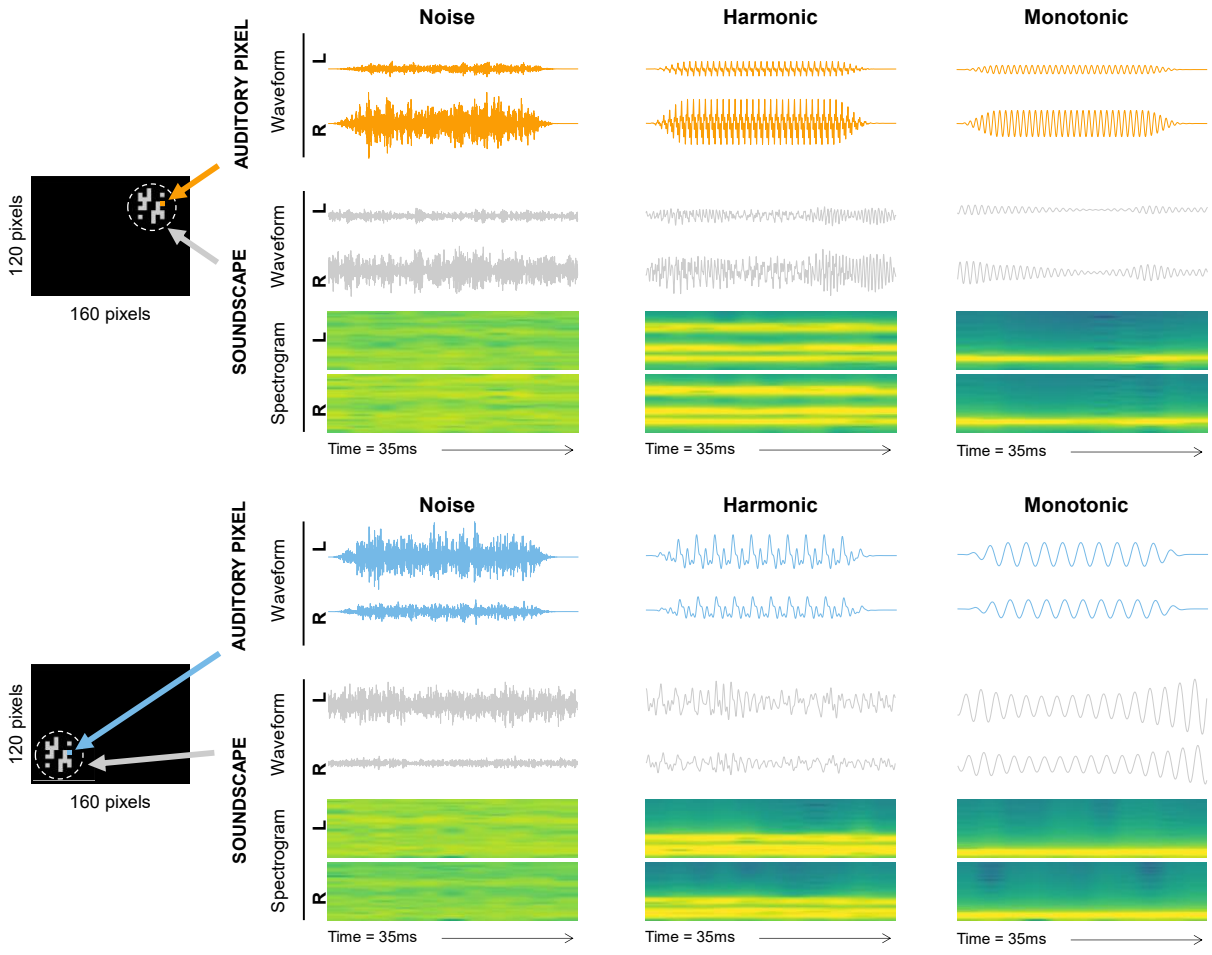
Video processing principles are similar to those used by Ambard et al. (2015), aiming to convey only new visual information from one frame to another. Video frames are grayscale images with gray levels ranging from 0 to 255. Pixels of the current frame are only conserved if the gray level pixel-by-pixel absolute difference with the previous frame (frame differencing) is larger than a threshold of 10. The processed image is then rescaled to a 160 × 120 (Horizontal × Vertical) grayscale image where 0-gray-level pixels are called "inactive" (i.e., no new visual information contained) and the others are "active" graphical pixels (i.e., containing new visual information). Active graphical pixels are then converted into spatialized sounds following a visual-to-auditory encoding in order to generate a soundscape (Figure 2), as explained in the following section.

2.2.4 Visual-to-auditory conversion

The visual-to-auditory conversion consists in the transformation of the processed video stream into a synchronized audio stream that acoustically encodes the extracted graphical features. Each graphical pixel is associated with an "auditory pixel" which is a stereophonic sound with auditory cues specific to the position of the graphical pixel it is associated with. The conversion from a graphical pixel to an auditory pixel follows an encoding that is explained step-by-step in the following sections. Each graphical pixel of a video frame is first associated with a corresponding monophonic audio pixel (detailed in Section 2.2.4.1). The spatialization of the sound using HRTFs

is then used to generate a stereophonic audio pixel that simulates a sound source with azimuth and elevation corresponding to the position of the graphical pixel in the virtual camera's field of view (detailed in Section 2.2.4.2). All the stereophonic pixels of a video frame are then compiled to obtain an audio frame (detailed in Section 2.2.4.3). Successive audio frames are then mixed together to generate a continuous audio stream (i.e., the soundscape). Two examples of stereophonic auditory pixels are provided in Figure 2 for each of the three encodings, as well as two examples of soundscapes depending on the location of an object in the field of view of the virtual camera.

Figure 2. Two examples of processed video frames and their corresponding soundscapes. The two processed video frames are depicted in the left side of the figure with a target located on the upper right (**top image**) and bottom left (**bottom image**). Active and inactive graphical pixels are depicted in gray and black, respectively. Two graphical pixels are highlighted in the video frames (orange in the top image, blue in the bottom image) and the corresponding auditory pixel waveforms are depicted in the right part of the figure in orange and blue. The corresponding soundscape waveforms (in gray) and soundscape spectrograms of the video frames are also depicted in the right part of the figure. Auditory pixel waveforms, soundscapes waveforms and spectrograms are displayed separately for the Noise encoding (left column), the Harmonic encoding (middle column) and the Monotonic encoding (right column) and for left (L) and right (R) ear channels separately.



2.2.4.1 Monophonic pixel synthesizing

Three visual-to-auditory encodings were tested in this study: the Noise encoding and 2 Pitch encodings (the Monotonic encoding and the Harmonic encoding). These methods varied in the elevation encoding scheme and in the spectral complexity of the monophonic auditory pixels but all three methods used afterwards the same method for the sound spatialization.

For the Noise encoding, the simulated sound source (i.e., monophonic auditory pixel) in the VAS was a white noise signal generated by inverting a Fourier representation of the auditory pixel with a flat spectrum and random phases.

For the Monotonic encoding, each monophonic auditory pixel was a sinusoidal waveform audio signal (i.e., a pure tone) with a random phase and a frequency related to the elevation of the corresponding graphical pixel in the processed image. For this purpose, we used a linear Mel scale ranging from 344 mel (bottom) to 1,286 mel (top) corresponding to frequencies from 250 to 1,492 Hz.

For the Harmonic encoding, we used the same monophonic auditory pixels as in the Monotonic encoding but instead of a pure tone, we added to it two other frequencies at the 2 following octaves with the same intensity and random phases.

Since the loudness depends on the frequency components of the audio signal, we minimized the differences in loudness between auditory pixels using the *pyloudnorm* Python-package (Steinmetz and Reiss, 2021). Auditory pixel spectrums were then adjusted to compensate for the frequency response of the headphones we used in this experiment (SONY MDR-7506).

2.2.4.2 Auditory pixel spatialization

The azimuth and elevation associated with each pixel were computed based on the coordinates of the corresponding graphical pixel in the camera's field of view. Monophonic auditory pixels were then spatialized by convolving them with the corresponding KEMAR HRTFs from the CIPIC database (Algazi et al., 2001a). This database provides HRTFs recordings with a sound source located in various azimuths and elevations ranging in steps of 5 and 5.625°, respectively. For each pixel, the applied HRTFs were estimated from the database by computing a 4 points time-domain interpolation in which the Interaural Level Difference (ILD) and the convolution signals were separately interpolated using bilinear interpolations before being reassembled as in Sodnik et al. (2005) but using a 2D interpolation instead of a 1D interpolation.

2.2.4.3 Audio frame mixing

Each auditory pixel lasted 34.83 ms including a 5 ms cosine fade-in and a 5 ms cosine fade-out. All the auditory pixels corresponding to the active graphical pixels of the processed current video frame were compiled to form an audio frame. After their compilation, these fade-in and fade-out were still present at the beginning and at the end of the audio frame and they were used to overlap successive audio frames while limiting the artifacts of the auditory transition.

2.2.5 Pointing tool and response collection

The pointing tool was a tracked gun pistol. Participants were instructed to indicate the perceived target position by pointing to it with the gun, with stretched arm. Participants logged their response by pressing a button with their index finger. They were instructed to hold the pointing tool with their dominant hand. The response position was defined as the intersection point of a virtual ray originating at the tip of the pointing tool and a virtual 1-m radius sphere with the origin at the location of the virtual camera. The response positions were declined in the elevation response and the azimuth response. The elevation and azimuth signed errors were also computed as the difference between the target position and the response position (in elevation and azimuth separately). A negative elevation signed error indicated a downward shift, and a negative azimuth signed error indicated a shift to the left in the response position. Unsigned errors were computed as the absolute value of the signed errors of each trial.

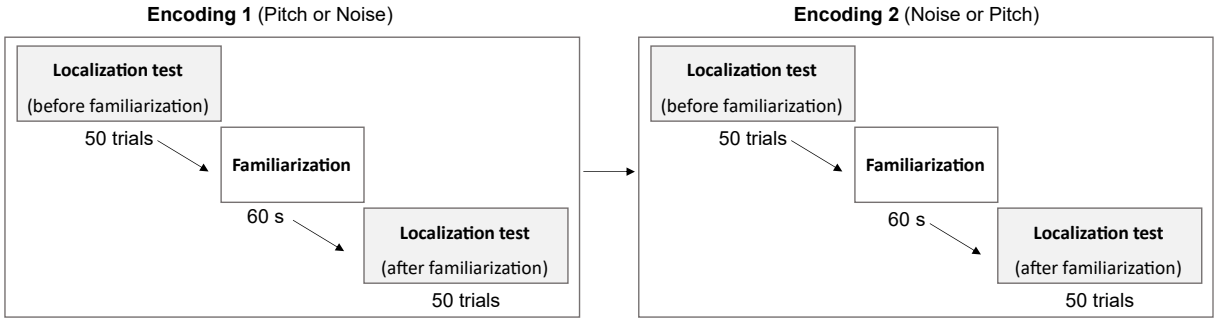
2.3 Experimental procedure

The experiment consisted in a 45-min session during which participants were seated comfortably in a chair at the center of a room surrounded by the virtual reality tracking system. The participants were equipped with SONY MDR-7506 headphones used to deliver soundscapes. Figure 3 illustrates the timeline of the experimental session. Each participant had to test two visual-to-auditory encodings: the Noise encoding, and a Pitch encoding (Monotonic or Harmonic encoding depending on the group they belonged). Participants from the Monotonic group had to test the Noise encoding and the Monotonic encoding, and participants from the Harmonic group had to test the Noise encoding and the Harmonic encoding. The order of the two tested encodings was counterbalanced between participants so half participants of each group started with the Noise encoding and the other half started with the Pitch encoding.

For each encoding, the participants practiced 2 times the localization test: one without any familiarization or explanation of the encoding and one after a familiarization session. At the beginning of the experiment, participants were instructed to localize a virtual target by pointing to it while being blindfolded. The experimenter explained that they will not be able to see the virtual

target, but that they will only hear it and that the sound will depend on the position of the target. No indication was given about the way visual-to-auditory encodings worked. Participants were seated and blindfolded using an opaque blindfold fixed with a rubber band and could remove it during breaks. Participants were instructed to keep their head as still as possible during the localization tests. For control purposes, participants' head position was recorded with the tracker every 200 ms to check that they kept their head still. We measured the maximum distance of the head from its mean position for each trial and we found an average maximum distance of approximately 1.5 cm showing that the instructions were rigorously followed.

Figure 3. Experimental timeline. Participants had to sequentially test the Noise encoding and one of the two Pitch encodings (Monotonic or Harmonic). Participants of the Monotonic and Harmonic groups tested the Monotonic encoding and the Harmonic encoding respectively. For each encoding, participants practiced the localization test two times, before and after a familiarization session.



2.3.1 Localization test

The localization test consisted in 50 trials during which blindfolded participants had to localize the virtual target using soundscapes provided by the visual-to-auditory SSD. During each localization test, the target was located at 25 different positions distributed on a grid of 5 azimuths (-40, -20, 0, +20, and +40°) and 5 elevations (-25, -12.5, 0, +12.5, and +25°). Figure 1B illustrates the grid with the 25 tested positions. As an example, the position [0°, 0°] corresponded to the central position, i.e., the virtual target was centered with the participant's head tracker. For the position [-40°, +12.5°], the target was 40° leftward and 12.5° upward from the central position ([0°, 0°]). The order of the tested positions was randomized and each position was tested 2 times per localization test. The target was placed at 1-meter-distance from the participant's head tracker for all positions (on the virtual 1-meter radius sphere used to collect the response positions).

Each trial started with a 500 ms 440 Hz beep sound, indicating the beginning of the trial. After a 500 ms silent period, the virtual target was displayed at one of the 25 tested positions. Participants were instructed to point with the pointing tool to the perceived location of the target with stretched arm. No time limit was imposed for responding but participants were asked to

respond as fast and accurately as possible. The virtual target was displayed until participants pressed the trigger of the pointing tool. The response position was recorded (see Section 2.2.5 for response position computing) and the target disappeared. After a 1,000 ms inter-trial break, the next trial began with the 500 ms beep sound. No feedback was provided regarding response accuracy.

2.3.2 Familiarization session

In between the 2 localization tests of each of the 2 tested encodings, participants practiced a familiarization session which consisted in a 60-s period during which participants freely moved the pointing tool in the front field. Figure 1C illustrates the familiarization session. The virtual target was continuously placed (i.e., no need to press the trigger) on a 1-meter radius sphere centered with the camera position, on the axis of the pointing tool. Consequently, when participants moved their arm, the target was continuously placed at the corresponding position on the 1-meter radius sphere and they could hear the soundscape provided by the encoding corresponding to the processed target images within the camera's field of view. The virtual camera position was updated one time at the beginning of the 60-s timer.

2.4 Data analysis

Statistical analysis were performed using R (version 3.6.1) (Team, 2020). Localization performance during localization tests was assessed separately for azimuth and elevation dimensions, with error-based and regression-based metrics, both fitted with Linear mixed models (LMMs) in order to take into account participants as random factor. All trials of all participants were included in the models without averaging the response positions or the unsigned errors by participant. The LMMs were fitted using the *lmerTest* R-package (Kuznetsova et al., 2017). We used an ANOVA with Satterthwaite approximation of degrees-of-freedom to estimate the effects. Post-hoc analysis were conducted using the *emmeans* R-package (version 1.7.4) (Lenth, 2022) with Tukey HSD correction.

2.4.1 Error-based metrics with unsigned and signed errors

Localization performance was assessed through unsigned and signed errors. The elevation signed errors and azimuth signed errors were computed as the difference between target position and response position in each trial. A negative elevation signed error indicated a downward shift, and a negative azimuth signed error indicated a shift to the left in the response. Only descriptive statistics were conducted on the signed errors. The unsigned errors were computed as the absolute value of the signed error for each trial. They were investigated using LMMs including Encoding (Noise or Pitch), Group (Monotonic or Harmonic) and Phase (Before or After the familiarization)

as fixed factors. Therefore, the positions of the target were not included as a factor in the LMMs of the unsigned error. Participants were considered as random effect in both models.

2.4.2 Regression-based metrics with response positions

LMMs were also used for the analysis of the response positions. LMMs included Encoding (Noise or Pitch), Group (Monotonic or Harmonic), Phase (Before or After the familiarization), and Target position as fixed effects. The target elevation only, and the target azimuth only, were included in the elevation response LMM, and in the azimuth response LMM, respectively. Participants were considered as random effect in both models. We used the LMMs predictions to approximate the elevation and the azimuth gains and biases. The gains and biases were obtained by computing the trends (slopes) and intercepts of the models expressing the response position as a function of target position. Note that an optimal localization performance would be obtained with a gain value of 1.0 and a bias of 0.0°.

3. Results

3.1 Performance in elevation localization

The elevation unsigned errors are depicted in Figure 4, left, all target positions combined. Table 1 shows the elevation signed and unsigned errors for each Target elevation, Phase, Encoding and Group. The ANOVA on elevation unsigned errors showed a significant interaction effect of Phase × Encoding × Group [$F(1, 7556) = 6.23$, $p = 0.0126$, $\eta_p^2 = 0.0008$]. Post-hoc analysis were conducted to investigate the interaction.

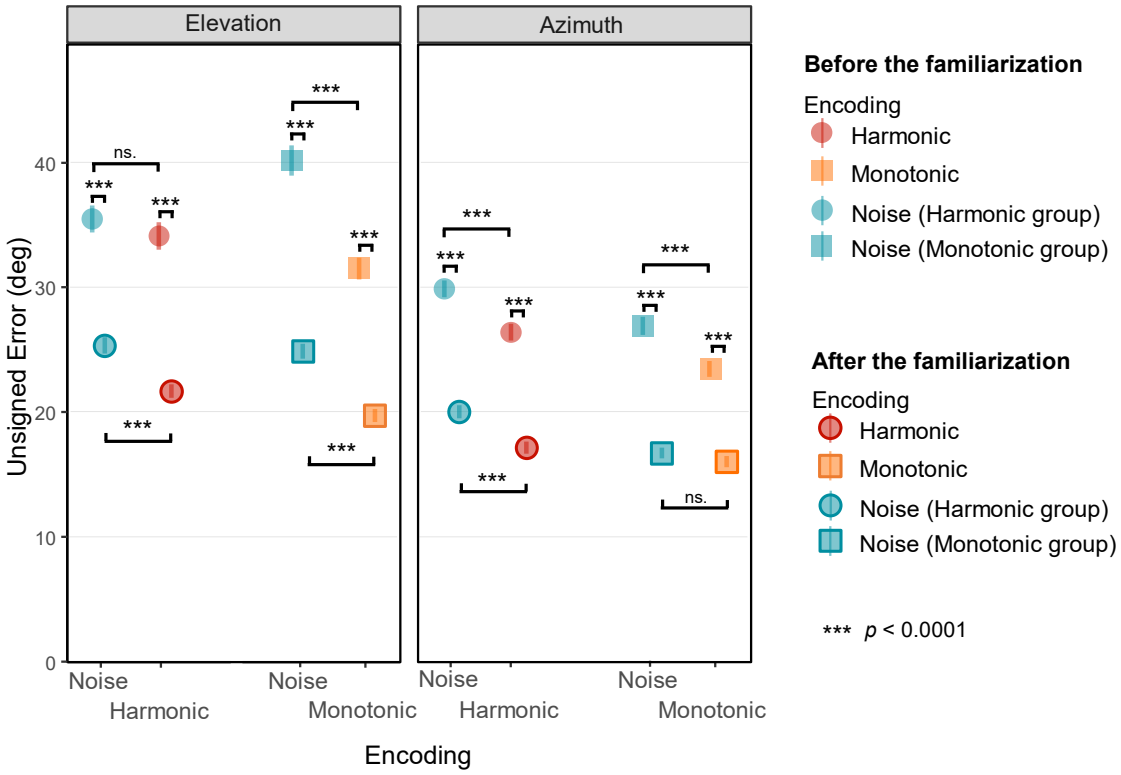
The elevation response positions are depicted in Figure 5. The ANOVA showed a significant interaction effect of Phase × Target Elevation × Encoding [$F(1, 7548) = 38.84$, $p < 0.0001$, $\eta_p^2 = 0.005$]. We conducted post-hoc analysis to investigate the elevation gain (the trend of the model) and bias (the intercept of the model) depending on the Phase and the Encoding. Although the interaction effect of Phase × Target Elevation × Encoding × Group was not significant [$F(1, 7548) = 0.50$, $p = 0.48$, $\eta_p^2 = 0.00007$], post-hoc analysis were also performed for a control purpose in order to check for differences between the Monotonic and Harmonic groups. The elevation response positions are provided separately for each participant in the Supplementary Figures S1, S2.

3.1.1 Elevation localization performance before the familiarization

Before the practice of the familiarization session, and depending on the encoding, the elevation unsigned errors were comprised between 31.54 ± 27.19° and 40.19 ± 37.02°. For the Monotonic group, the elevation unsigned errors were significantly lower with the Monotonic

encoding (M = 31.54, SD = 27.19) than with the Noise encoding (M = 40.19, SD = 37.03) [$t(7556) = 7.457, p < 0.0001$], suggesting a lower accuracy with the Noise encoding. There was no significant difference in the Harmonic group regarding the elevation unsigned error between the Harmonic encoding (M = 34.14, SD = 33.69) and the Noise encoding (M = 35.51, SD = 33.69).

Figure 4. Unsigned error in elevation (left) and in azimuth (right) as a function of the encoding, all target positions combined. Mean unsigned errors (in degree) before (non-surrounded) and after (surrounded) are depicted separately for the Monotonic group (squares) and Harmonic group (circles) and for the three visual-to-auditory encodings: the Noise (blue), the Monotonic (orange) and the Harmonic (red) encodings. Error bars show standard error of the unsigned error.

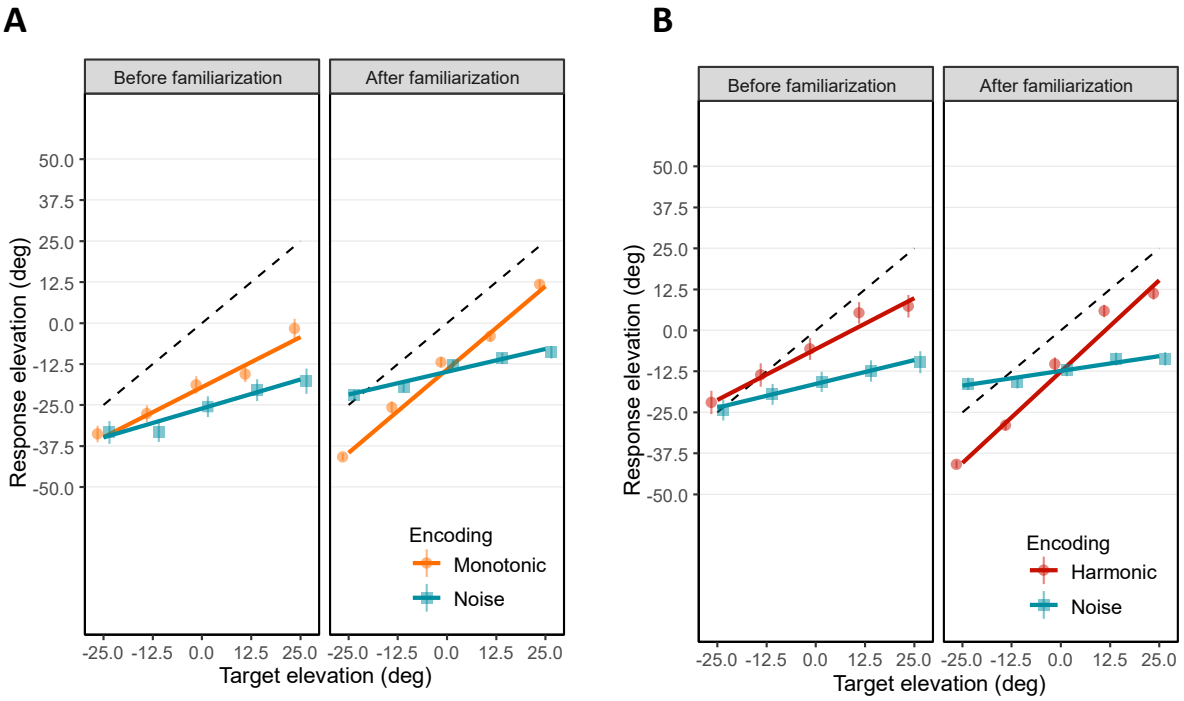


The elevation response positions before the familiarization are depicted in the left panels of the Figure 5A, B for the Monotonic group and the Harmonic group, respectively. The elevation gains were significantly different from 0.0 for all encodings: 0.62 [95% CI = [0.5, 0.74], $t(7548) = 10.118, p < 0.0001$] with the Harmonic encoding, 0.61 [95% CI = [0.49, 0.73], $t(7548) = 9.94, p < 0.0001$] with the Monotonic encoding, and 0.29 [95% CI = [0.17, 0.41], $t(7548) = 4.728, p < 0.0001$] and 0.35 [95% CI = [0.23, 0.47], $t(7548) = 5.746, p < 0.0001$] with the Noise encoding of the Harmonic and Monotonic groups, respectively. It suggests that participants could discriminate different elevation positions with the three encodings even before the familiarization.

However, elevation gains were significantly lower than the optimal gain 1.0 with all encodings: with the Harmonic encoding [$t(7548) = -6.173, p < 0.0001$], with the Monotonic encoding [$t(7548) = -6.351, p < 0.0001$], and with the Noise encoding of the Harmonic group

[t(7548) = -11.562, p < 0.0001] and of the Monotonic group [t(7548) = -10.544, p < 0.0001]. It depicts a situation where although some variations in elevation seemed to be perceived with the three encodings, participants had difficulties to estimate it before the familiarization.

Figure 5. Elevation response position as a function of target elevation in the Monotonic group (**A**) and the Harmonic group (**B**). Mean elevation response positions (in degree) before (left) and after (right) are represented separately for the three visual-to-auditory encodings: the Noise (blue squares), the Monotonic (orange circles) and the Harmonic (red circles) encodings. Error bars show standard error of elevation response position. Solid lines represent the elevation gains with the Noise (blue), the Monotonic (orange) and Harmonic (red) encodings. Black dashed lines indicate the optimal elevation gain 1.0.



The participants tended to localize the elevation with a higher performance with the Harmonic or Monotonic encoding than with the Noise encoding. Indeed, the participants from the Harmonic group showed a higher elevation gain with the Harmonic encoding than with the Noise encoding with a significant difference of 0.33 [t(7548) = -3.811, p = 0.0008]. For the Monotonic group, the elevation gain was also significantly higher with the Monotonic encoding than with the Noise encoding with a difference of 0.26 [t(7548) = -2.97, p = 0.016]. There was no significant difference regarding the elevation gain between the Harmonic and the Monotonic encodings.

The participants tended to underestimate the elevation position of the targets with the three encodings, as indicated by downward bias and negative elevation errors. In the Monotonic group, the elevation bias were -26.02° (95% CI = [-31.9, -20.16]) with the Noise encoding and -19.52° (95% CI = [-25.4, -13.65]) with the Monotonic encoding. In the Harmonic group, the elevation

bias with the Noise encoding and with the Harmonic encoding were -16.33° (95% CI = [-22.2, -10.47]) and -5.73° (95% CI = [-11.6, 0.14]), respectively. With the exception of the Harmonic encoding for which there was just a trend [$t(44.9) = 1.97$, $p = 0.055$], all the elevation bias mentioned above were significantly negative [all $|t(44.9)| > 5.61$, all $p < 0.0001$].

To sum up, participants appeared partially able to perceive a variation of the elevation position of the target with the three encodings before the audio-motor familiarization. Interestingly, participants seemed better able to localize the elevation with the Harmonic and Monotonic encodings.

Table 1. Elevation signed error and unsigned error (in degree) for each encoding and target elevation, before, and after the familiarization session.

Encoding	Target elevation (degree)	Elevation signed error (degree) Mean ± standard deviation		Elevation unsigned error (degree) Mean ± standard deviation	
		Before familiarization	After familiarization	Before familiarization	After familiarization
Monotonic	+25	-26.71 ± 41.40	-13.21 ± 20.02	37.70 ± 31.55	18.55 ± 15.17
	+12.5	-28.09 ± 33.09	-16.55 ± 22.28	33.45 ± 27.62	21.99 ± 16.89
	0	-18.83 ± 36.41	-11.94 ± 22.73	31.63 ± 26.01	19.60 ± 16.55
	-12.5	-15.14 ± 34.95	-13.23 ± 24.95	27.30 ± 26.50	20.73 ± 19.14
	-25	-8.82 ± 34.34	-15.81 ± 15.18	27.51 ± 22.29	17.89 ± 21.65
Harmonic	+25	-17.66 ± 47.39	-13.73 ± 26.16	36.66 ± 34.76	23.03 ± 18.45
	+12.5	-7.18 ± 45.40	-6.61 ± 26.46	33.41 ± 31.47	20.73 ± 17.67
	0	-5.68 ± 45.71	-10.28 ± 28.07	32.91 ± 32.14	24.06 ± 17.67
	-12.5	-1.10 ± 48.75	-16.43 ± 21.80	33.44 ± 35.40	22.23 ± 15.81
	-25	2.99 ± 48.76	-15.85 ± 16.08	34.28 ± 34.72	18.44 ± 13.01
Noise (Monotonic group)	+25	-42.69 ± 52.91	-33.92 ± 27.00	56.49 ± 37.73	37.54 ± 21.64
	+12.5	-32.93 ± 45.45	-23.24 ± 23.54	43.63 ± 35.23	28.00 ± 17.56
	0	-25.49 ± 43.51	-12.98 ± 24.78	36.61 ± 34.62	23.09 ± 15.73
	-12.5	-20.61 ± 43.35	-7.09 ± 22.23	32.93 ± 34.88	19.02 ± 13.46
	-25	-8.40 ± 47.90	3.11 ± 22.32	31.30 ± 37.15	16.86 ± 14.90
Noise (Harmonic group)	+25	-34.74 ± 46.47	-33.67 ± 28.35	47.86 ± 32.71	36.96 ± 23.87
	+12.5	-24.89 ± 45.12	-21.25 ± 27.08	40.59 ± 31.66	27.87 ± 20.16
	0	-15.68 ± 42.71	-12.20 ± 25.02	32.41 ± 31.87	22.15 ± 16.81
	-12.5	-7.03 ± 44.28	-3.40 ± 25.66	28.84 ± 34.27	19.68 ± 16.76
	-25	0.69 ± 43.84	8.72 ± 25.46	27.84 ± 33.81	20.10 ± 17.85

3.1.2 Elevation localization performance after the familiarization

After the familiarization, the elevation unsigned errors were significantly higher with the Noise encoding than with the 2 pitch-based encodings (Monotonic or Harmonic encodings). With the Noise encoding, the elevation unsigned errors were 24.90 ± 18.40° in the Monotonic group and 25.35 ± 20.31° in the Harmonic group. With the Harmonic and Monotonic encodings, the elevation unsigned errors were 21.70 ± 16.72° and 19.75 ± 16.25° respectively. In the Monotonic group, the elevation unsigned errors were significantly lower with the Monotonic encoding (M = 19.75, SD = 16.25) than with the Noise encoding (M = 24.90, SD = 18.40) [$t(7556) = 4.44$, $p < 0.0001$]. Unlike before the familiarization, the difference was also significant in the Harmonic group. The elevation unsigned errors with the Harmonic encoding (M = 21.70, SD = 16.72) were lower than with the Noise encoding (M = 25.35, SD = 20.31), [$t(7556) = 3.15$, $p = 0.0016$]. Interestingly, the elevation unsigned errors significantly decreased after the familiarization with all the encodings [all $|t(7556)| > 8.75$, all $p < 0.0001$], suggesting that participants localized more accurately the elevation after the familiarization.

The elevation response positions after the familiarization are depicted in the Figure 5A, B for the Monotonic and Harmonic groups, respectively. After the familiarization, the elevation gains were still significantly higher than 0.0 [all $|t(7548)| > 2.9152$, all $p < 0.0036$] with all encodings in the 2 groups. The elevation gains were 1.112 (95% CI = [0.99, 1.23]) with the Harmonic encoding and 1.015 (95% CI = [0.89, 1.14]) with the Monotonic encoding. For the participants of the Harmonic group and the Monotonic group, the elevation gains with the Noise encoding were 0.179 (95% CI = [0.06, 0.30]), and 0.278 (95% CI = [0.16, 0.40]), respectively.

The elevation gains were significantly higher with the Harmonic and Monotonic encodings than with the Noise encoding. We measured a difference of 0.74 [$t(7548) = -8.49$, $p < 0.0001$] in the Harmonic group and a difference of 0.93 [$t(7548) = -10.75$, $p < 0.0001$] in the Monotonic group. Inter-group analysis showed that the difference in elevation gain between the Monotonic and the Harmonic encodings did not significantly differ [$t(7548) = 1.12$, $p = 0.95$].

The elevation gains with the Harmonic and the Monotonic encodings significantly improved after the familiarization to get closer than the optimal gain 1.0. With the Harmonic encoding, the elevation gain significantly increased from 0.62 to 1.112 [$t(7548) = 5.66$, $p < 0.0001$] after which it was not significantly different from the optimal gain 1.0 [$t(7548) = 1.832$, $p = 0.067$]. With the Monotonic encoding, the elevation gain significantly increased from 0.61 to 1.015 [$t(7548) = 4.665$, $p < 0.0001$], and was also no more significantly different from the optimal gain 1.0 [$t(7548) = 0.246$, $p = 0.806$]. However, with the Noise encoding in both groups, the familiarization did not improve the elevation gains. In the Harmonic and Monotonic groups, the elevation gains decreased

from 0.29 to 0.179 and from 0.35 to 0.278, respectively, but as previously reported, the decreases were not significant.

Participants kept tending to underestimate the elevation position of the targets with all three encodings, as indicated by persistent negative bias. In the Monotonic group, the elevation bias with the Noise encoding and with the Monotonic encoding were -14.82° (95% CI = [-20.7, -8.96]) and -14.15° (95% CI = [-20.0, -8.28]), respectively. In the Harmonic group, the elevation bias with the Noise encoding and with the Harmonic encoding were -12.36° (95% CI = [-18.2, -6.49]) and -12.58° (95% CI = [-18.4, -6.72]), respectively. All the elevation bias were significantly negative [all $|t(44.9)| > 4.24$, all $p < 0.0001$].

To sum up, after the familiarization, the perception of elevation with the Harmonic and Monotonic encodings improved with elevation gains getting closer to the optimal gain. However, the familiarization did not induce any significant improvement in the perception of elevation with the Noise encoding, with persistent low elevation gains in both groups. Additionally, the underestimation elevation bias decreased with the Monotonic and Noise encodings, but not with the Harmonic encoding for which it increased.

3.2 Performance in azimuth localization

The azimuth unsigned errors are depicted in Figure 4, right, all target positions combined. Table 2 shows the azimuth signed and unsigned errors for each Target azimuth, Phase, Encoding and Group. The ANOVA on azimuth unsigned errors showed a significant interaction effect of Phase × Encoding [$F(1, 7556) = 5.15$, $p = 0.023$, $\eta_p^2 = 0.00068$], but the interaction including the group was not significant.

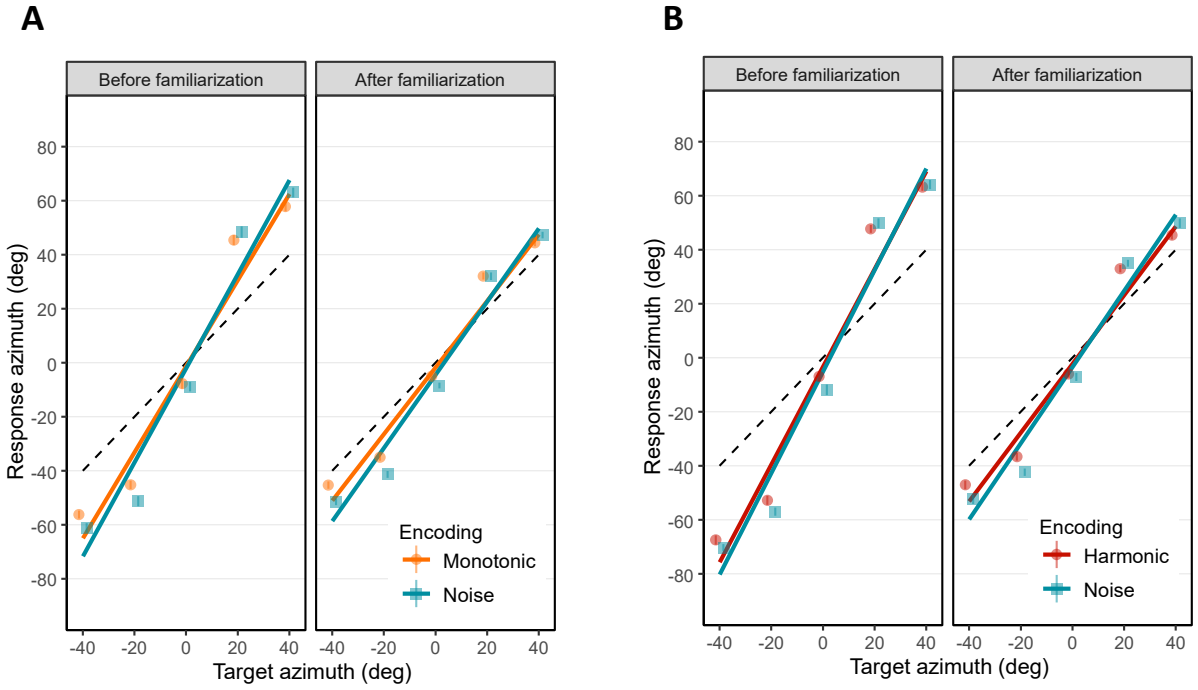
The azimuth response positions are depicted in Figure 6. The ANOVA yielded a significant interaction effect of Phase × Target Azimuth × Encoding [$F(1, 7548) = 12.69$, $p = 0.0004$, $\eta_p^2 = 0.00005$]. Post-hoc analysis were conducted to investigate the azimuth gain (the trend of the model) and bias (the intercept of the model) depending on the Phase and the Encoding. Although the interaction effect of Phase × Target Elevation × Encoding × Group was not significant [$F(1, 7548) = 1.64$, $p = 0.20$, $\eta_p^2 = 0.0002$], we conducted post-hoc analysis to check for differences between the Monotonic and Harmonic groups for a control purpose. The azimuth response positions are provided separately for each participant in the Supplementary Figures S3, S4.

3.2.1 Azimuth localization performance before the familiarization

Before the practice of the familiarization session, and depending on the encoding, the azimuth unsigned errors were comprised between 23.48 ± 19.48° and 29.91 ± 20.38°. In the

Monotonic group, the azimuth unsigned errors were significantly lower with the Monotonic encoding (M = 23.48, SD = 19.48) than with the Noise encoding (M = 26.92, SD = 22.58) [$t(7556) = 4.64$, $p < 0.0001$]. The azimuth unsigned errors in the Harmonic group were also significantly lower [$t(7556) = 4.73$, $p < 0.0001$] with the Harmonic encoding (M = 26.41, SD = 20.63) than with the Noise encoding (M = 29.91, SD = 33.69).

Figure 6. Azimuth response position as a function of target azimuth in the Monotonic group (A) and the Harmonic group (B). Mean azimuth response positions (in degree) before (left) and after (right) are represented separately for the three visual-to-auditory encodings: the Noise (blue squares), the Monotonic (orange circles) and the Harmonic (red circles) encodings. Error bars show standard error of azimuth response position. Solid lines represent the azimuth gains with the Noise (blue), the Monotonic (orange) and Harmonic (red) encodings. Black dashed lines indicate the optimal azimuth gain 1.0.



The azimuth response positions over all participants before the familiarization are depicted in the left panels of the Figure 6 A, B for the Monotonic and Harmonic groups, respectively. Before the familiarization, the participants were able to interpret soundscapes to localize the target azimuth. First, the participants perceived different azimuth positions. Indeed, azimuth gains were significantly different from 0.0 with all encodings: 1.81 (95% CI = [1.75, 1.87], $t(7548) = 70.397$, $p < 0.0001$) with the Harmonic encoding, 1.59 [95% CI = [1.54, 1.65], $t(7548) = 61.996$, $p < 0.0001$] with the Monotonic encoding, and 1.88 [95% CI = [1.82, 1.93], $t(7548) = 73.0624$, $p < 0.0001$] and 1.74 [95% CI = [1.68, 1.80], $t(7548) = 67.710$, $p < 0.0001$] with the Noise encoding for the participants in the Harmonic and Monotonic groups, respectively.

Table 2. Azimuth signed error and unsigned error (in degree) for each encoding and target azimuth, before, and after the familiarization session.

Encoding	Target Azimuth (degree)	Azimuth signed error (degree) Mean ± standard deviation		Azimuth unsigned error (degree) Mean ± standard deviation	
		Before familiarization	After familiarization	Before familiarization	After familiarization
Monotonic	+40	17.79 ± 23.73	4.39 ± 18.25	23.16 ± 18.5	13.97 ± 12.5
	+20	25.34 ± 25.50	12.06 ± 18.35	28.71 ± 21.61	17.14 ± 13.69
	0	-7.75 ± 25.39	-5.02 ± 19.33	17.95 ± 19.52	15.21 ± 12.90
	-20	-25.25 ± 21.81	-15.02 ± 18.10	26.93 ± 19.69	18.68 ± 14.27
	-40	-16.27 ± 20.33	-5.34 ± 19.40	20.68 ± 15.79	15.23 ± 13.11
Harmonic	+40	23.26 ± 19.59	5.43 ± 21.49	24.9 ± 17.45	15.83 ± 15.48
	+20	27.17 ± 20.26	12.98 ± 18.80	28.17 ± 19.61	17.23 ± 14.98
	0	-6.89 ± 23.03	-6.02 ± 18.49	16.89 ± 17.36	14.15 ± 12.97
	-20	-32.76 ± 23.38	-16.61 ± 19.93	33.16 ± 22.81	21.6 ± 14.35
	-40	-27.45 ± 23.93	-7.03 ± 22.08	29.25 ± 21.68	16.68 ± 16.05
Noise (Monotonic group)	+40	23.01 ± 25.34	7.27 ± 16.58	27.91 ± 19.79	14.35 ± 11.01
	+20	28.35 ± 29.17	11.96 ± 19.59	30.95 ± 26.38	18.11 ± 14.06
	0	-8.89 ± 24.82	-8.47 ± 14.79	16.78 ± 20.31	12.04 ± 12.05
	-20	-31.27 ± 25.61	-21.37 ± 17.24	33.42 ± 22.71	22.32 ± 15.98
	-40	-21.22 ± 24.10	-11.55 ± 16.88	25.53 ± 19.45	16.82 ± 11.59
Noise (Harmonic group)	+40	24.13 ± 23.85	9.86 ± 24.77	28.65 ± 18.12	19.17 ± 18.49
	+20	29.80 ± 24.77	15.02 ± 17.67	31.78 ± 22.16	18.94 ± 13.35
	0	-11.84 ± 27.02	-7.04 ± 20.76	21.10 ± 20.57	16.34 ± 14.57
	-20	-36.95 ± 20.92	-22.36 ± 18.41	37.28 ± 20.31	25.06 ± 14.48
	-40	-30.29 ± 17.83	-12.40 ± 23.27	30.71 ± 17.09	20.71 ± 16.28

Interestingly, the azimuth gains were significantly higher than the optimal gain (i.e., higher than 1.0) with the Harmonic encoding [$t(7548) = 31.492$, $p < 0.0001$], the Monotonic encoding [$t(7548) = 23.091$, $p < 0.0001$], and the Noise encoding in the Harmonic group [$t(7548) = 34.156$, $p < 0.0001$] and the Monotonic group [$t(7548) = 28.804$, $p < 0.0001$]. These gains higher than the optimal gain reflect a lateral overestimation (i.e., left targets localized too much on the left and right targets localized too much on the right) that can be seen with the three encodings.

In the Monotonic group, the overestimation observed with the Noise encoding was significantly higher than with the Monotonic encoding [$t(7548) = 4.04$, $p = 0.0003$]. In the Harmonic group, the overestimation with the Noise encoding compared to the Harmonic encoding was also higher but not significantly. Inter-group comparison of the azimuth gains obtained with the Noise encoding shows a small but significant higher azimuth gain in the Harmonic group [difference of 0.14: $t(7548) = 3.784$, $p = 0.0009$]. As inter-group comparison, we also observed a

slight but significant higher overestimation pattern with the Harmonic encoding in comparison with the Monotonic encoding [difference of 0.22: $t(7548) = 5.94$, $p < 0.0001$].

Another interesting result is the tendency to show a left shift as indicated by negative azimuth bias with the three encodings. With the Noise encoding in the Harmonic group, the leftward azimuth bias of -5.03° was significant [$t(47.2) = 2.84$, $p = 0.0066$]. However, leftward azimuth bias with the other encodings were not significantly different from 0.0° (Harmonic group, Noise encoding: -3.23°; Monotonic group, Noise encoding: -2.0°; Monotonic group, Monotonic encoding: -1.233°).

In summary, before the familiarization and with the three encodings, participants were able to localize the azimuth of the target accurately with a tendency to overestimate the lateral eccentricity and a tendency to point too much on the left.

3.2.2 Azimuth localization performance after the familiarization

After the participants practiced the familiarization session, and depending on the encoding, the azimuth unsigned errors were comprised between 16.05 ± 13.38° and 20.05 ± 15.77°. In the Harmonic group, the azimuth unsigned errors were significantly lower with the Harmonic encoding (M = 17.16, SD = 14.97) than with the Noise encoding (M = 20.05, SD = 15.77) [$t(7556) = 3.91$, $p = 0.0001$]. The azimuth unsigned errors were not significantly different anymore between the Monotonic encoding (M = 16.05, SD = 13.38) and the Noise encoding (M = 16.73, SD = 13.50). Importantly, the azimuth unsigned errors significantly decreased after the familiarization session for all three encodings [all $|t(7556)| > 10.06$, all $p < 0.0001$], suggesting that participants localized more accurately the azimuth after the familiarization.

The azimuth response positions after the familiarization are depicted in the right panels of the Figure 6A, B for the Monotonic and Harmonic groups, respectively. As expected, after the familiarization, participants were still able to localize different azimuth positions by interpreting soundscapes. Azimuth gains were still significantly different from 0.0 with the Harmonic encoding [1.27, 95% CI = [1.22, 1.32], $t(7548) = 49.51$, $p < 0.0001$], with the Monotonic encoding [1.23, 95% CI = [1.18, 1.28], $t(7548) = 47.96$, $p < 0.0001$], and with the Noise encoding for the participants in the Harmonic group [1.41, 95% CI = [1.36, 1.46], $t(7548) = 54.84$, $p < 0.0001$] and in the Monotonic group [1.35, 95% CI = [1.30, 1.41], $t(7548) = 52.711$, $p < 0.0001$], respectively.

The overestimation pattern was still present, as indicated by azimuth gains still significantly higher than the optimal gain 1.0 with all encodings: the Harmonic encoding [$t(7548) = 10.61$, $p < 0.0001$], the Monotonic encoding [$t(7548) = 9.06$, $p < 0.0001$], the Noise encoding in the Harmonic group [$t(7548) = 15.93$, $p < 0.0001$] and in the Monotonic group [$t(7548) = 13.81$, $p < 0.0001$].

Although the lateral overestimation was still significant, it significantly decreased compared to the same localization test before the familiarization. Indeed, the azimuth gains decreased and reached values closer than the optimal gain 1.0 with the 3 encodings. There were significant decreases in azimuth gains of a magnitude of 0.54 [$t(7548) = 14.77$, $p < 0.0001$] and 0.36 [$t(7548) = 9.92$, $p < 0.0001$] with the Harmonic and Monotonic encodings, respectively. The decreases in azimuth gains with the Noise encoding in the Harmonic and the Monotonic groups were also significant with a decrease of a magnitude of, respectively, 0.47 [$t(7548) = 12.897$, $p < 0.0001$] and 0.39 [$t(7548) = 10.61$, $p < 0.0001$].

Additionally, after the familiarization, participants tended to localize the azimuth with a higher performance with the Harmonic and Monotonic encodings than with the Noise encoding. This is suggested by a more pronounced lateral overestimation with the Noise encoding in both groups: the azimuth gains were 0.14 higher [$t(7548) = 3.76$, $p = 0.0042$] and 0.12 higher [$t(7548) = 3.36$, $p = 0.018$] with the Noise encoding in comparison with the Harmonic and Monotonic encodings, respectively.

The slight tendency to show a left shift bias in azimuth was still present with the three encodings. With the Noise encoding in the Monotonic group, the leftward azimuth bias of -4.43° was significant [$t(47.2) = 2.502$, $p = 0.0159$], but in the Harmonic group the bias of -3.38° was just a tendency [$t(47.2) = 1.911$, $p = 0.0621$]. The leftward azimuth bias with the Harmonic encoding (-2.25°) and Monotonic encoding (-1.79°) were also not significant.

To sum up the accuracy in azimuth localization, participants were able to localize target azimuths accurately even before the audio-motor familiarization. After the familiarization, the accuracy increased with a decrease in both the tendency to overestimate the lateral position of lateral targets and the tendency to point too much on the left.

4. Discussion

In this study, we investigated the early stage of use of visual-to-auditory SSDs based on the creation of a VAS (Virtual Acoustic Space) for object localization in a virtual environment. Based on soundscapes created using non-individualized HRTFs, we investigated blindfolded participants' abilities to localize a virtual target with three encoding schemes: one conveying elevation with spatialization only (Noise encoding), and two conveying elevation with spatialization and pitch modulation (Monotonic and Harmonic encodings). The two pitch-based encodings varied regarding the sound spectrum complexity: one narrowband with monotones (Monotonic encoding) and one more complex with 2 additional octaves (Harmonic encoding). In order to compare the localization abilities for the azimuth and the elevation with the different visual-to-

auditory encodings, we collected the response positions and angular errors of the participants during a task consisting in the localization of a virtual target placed at different azimuths and elevations in their front-field.

4.1 Elevation localization abilities using the visual-to-auditory encodings

4.1.1 Elevation localization performance only based on non-individualized HRTFs is impaired

With the spatialization-based only encoding (Noise encoding), the target was localized before the familiarization with an elevation unsigned error between 27.84 ± 33.81° and 56.49 ± 37.73°. After the familiarization, the elevation unsigned errors decreased to reach values comprised between 16.86 ± 14.90° and 37.54 ± 21.64°. As a comparison, in Mendonça et al. (2013) where the same HRTFs database was used with a white noise sound, the mean elevation unsigned error of participants was 29.3° before practicing a training. The elevation unsigned errors in Geronazzo et al. (2018) without any familiarization and with a white noise sound were comprised between 15.58 ± 12.47° and 33.75 ± 16.17° depending on participants, which is comparable to our results after the familiarization. However, as shown by elevation gains below 0.4 before or after familiarization, the participants had difficulties to discriminate different elevations with this encoding.

The abilities to localize the elevation of an artificially spatialized sound are known to be impaired in comparison with azimuth (Wenzel et al., 1993). Those difficulties arise from the spectral distortions that are specific to individual body morphology (Blauert, 1996; Xu et al., 2007). When using non-individualized HRTFs, these spectral distortions are different from the participant's specific distortions, causing misinterpretation of elevation location. Additionally, the abilities to localize the elevation position of a sound source (virtual or real) are modulated by the spectral content of the sound (Middlebrooks and Green, 1991; Blauert, 1996).

In our study, the difficulty with the spatialization-based only encoding to localize the elevation of the target, even after the audio-motor familiarization, could be explained by a too brief training period to get used to the new auditory cues. Actually, some studies showed an improvement of localization abilities with non-individualized or modified HRTFs after 3 weeks of training in Majdak et al. (2013) or Romigh et al. (2017), or after 2 weeks in Shinn-Cunningham et al. (1998) or 1 week in Kumpik et al. (2010), and about 5 h in Bauer et al. (1966). Moreover, Mendonça et al. (2013) showed the positive long term effect (1-month long) of training in azimuth and elevation localization abilities with a sound source spatialized using the same HRTFs database that was used in the current study. It suggests that the exclusive use of HRTFs to encode spatial

information in SSDs might require a long training period or a long process to acquire individualized HRTFs.

4.1.2 Positive effects of cross-modal correspondence on elevation localization

The participants' abilities to localize the elevation of the target using the 2 pitch-based encodings were significantly better than with a broadband sound spatialization encoding. Before the audio-motor familiarization, with the narrowband encoding (Monotonic) and the more complex encoding (Harmonic), the unsigned errors in elevation were comprised between 27.30 ± 26.50° and 37.79 ± 31.55° depending on the target elevation.

Before the familiarization, participants did not receive any information about the way the sound was modulated depending on the target location. In other words, they did not know that low pitch sounds were associated with low elevation locations, and conversely. However, the individual results of each participant for the elevation (Supplementary Figures S1, S2) suggest that even before the familiarization, several participants interpreted the pitch to perceive the target elevation, using high pitch for high elevation and low pitch for low elevation. We suppose that participants were able to guess that the pitch of the sound varied with the target elevation because the experimenter explicitly told them that sound features were modulated as a function of the location of the target although no details regarding this modulation were provided. Two participants (S12 from the Harmonic group and S15 from the Monotonic group) reversed the pitch encoding by associating a low pitch to high elevations and a high pitch to low elevations, but they reversed this miss-representation after the familiarization. Our study showed that after the audio-motor familiarization, the elevation unsigned errors significantly decreased with both pitch-based encodings to reach values comprised between 17.67 ± 22.23° and 24.06 ± 17.67°, which are lower elevation unsigned errors than the mean elevation error of 25.2° immediately after the training in Mendonça et al. (2013).

In the visual-to-auditory SSD domain, the artificial pitch mapping of elevation is used by several existing visual-to-auditory SSDs and relies on the audiovisual cross-modal correspondence between visual elevation and pitch (Spence, 2011; Deroy et al., 2018). In the current study, the frequency range was between 250 Hz and about 1,500 Hz with the Monotonic encoding and between 250 Hz and about 6,000 Hz with the Harmonic encoding (i.e., 1,500 Hz × 2 × 2). The floor value of 250 Hz was chosen to provide frequency steps of at least 3 Hz between each of the 120 auditory pixels in a column, to fit to the human frequency discrimination abilities (Howard and Angus, 2009). We used the Mel scale (Stevens et al., 1937) to take into account the perceived scaling in sound frequency discrimination. All the SSDs using a pitch mapping of elevation use different frequency ranges, resolutions (i.e., number of used frequencies) and frequency steps. The vOICe

SSD (Meijer, 1992) uses a larger frequency range than the current study (from 500 to 5,000 Hz) following an exponential scale with a 64-frequency resolution. The EyeMusic SSD (Abboud et al., 2014) uses a pentatonic musical scale with 24 frequencies from 65.785 Hz to 1577.065 Hz. The SSD proposed in Ambard et al. (2015) also uses 120 frequency steps but following the Bark scale (Zwicker, 1961) and with a larger frequency range (from 250 Hz to about 2,500 Hz). Technically, increasing the range of frequencies might increase discrimination abilities between target elevations and improve localization abilities. Although, as sound frequency increases the sound feels unpleasant (Kumar et al., 2008). We can postulate that SSD users should be able to modulate some of the parameters in order to adapt the encoding scheme to their own auditory abilities and perceptual preferences.

Our results suggest that a pitch mapping of elevation can quickly be interpreted, even without any explicit explanation of the mapping rules. They also suggest that the pitch mapping provides acoustic cues that are easily interpretable at the early stage of use of a SSD to localize an object. In terms of spatial perception, our study shows that adding abstract acoustic cues to convey spatial information can be more efficient than an imperfect synthesizing of natural acoustic cues. It is difficult to assert that the differences in the results between the Noise encoding and the Pitch encodings are entirely due to the cross-modal correspondence between elevation and pitch since modifying the timbre of the sound by reducing its spectral content also modified how the HRTFs spatialize the sound. Therefore, it would be interesting to investigate the localization performance with monotonic or harmonic sounds in which the pitch is constant (i.e., not related to the elevation of the target) and by conducting an experiment where HRTFs convolution is computed to convey azimuth only, with for instance a constant elevation of 0°.

4.1.3 Insights about the pitch-elevation cross-modal correspondence

Although the aim of this study was not to directly investigate the multisensory perceptual process, the results might bring insights about the pitch-elevation cross-modal correspondence. In the SSD research, it has been suggested that the pitch-based elevation mapping is intuitive in an object recognition task (Stiles and Shimojo, 2015). Based on the results of the current study, it also seems intuitive in a localization task. However, it remains to be further investigated with, for instance, a comparison of elevation localization abilities with a similar pitch-based elevation encoding and another encoding where the direction of the pitch mapping is reversed (i.e., low pitch for high elevation and high pitch for low elevation). The current study also raises the question regarding the automaticity of the cross-modal correspondences as disccussed in Spence and Deroy (2013). In the current study, the facilitation effect of the cross-modal correspondence probably relies on voluntary multisensory perceptual processes. The way the instructions were given to the

participants intrinsically induced a goal-directed voluntary strategy in order to infer which modifications in the sound could convey information about the location of the object.

These insights about multisensory process should also be investigated in the blind. Since the pitch-elevation cross-modal correspondence has been suggested to be weak in this population (Deroy et al., 2016), and since auditory spatial perception of the elevation can be impaired in this population (Voss, 2016), it remains to investigate whether similar results would be obtained with blind participants. For this reason, the procedure of the current study was designed in a way to be reproducible with blind participants.

4.1.4 No positive effects of harmonics on elevation localization

The elevation-pitch encoding adds a salient auditory cue while reducing the frequency range where the HRTFs spectrum alterations can operate. To study the effect of the spectral complexity we used an encoding with harmonic sounds (monotonic and 2 following octaves) meant to be a trade-off in terms of spectral complexity between the broadband sound of the Noise encoding and the monotones of the Monotonic encoding. Although pure tones were used in the Monotonic encoding, it is important to keep in mind that soundscapes were not pure tones. Indeed, soundscapes were made of adjacent auditory pixels, resulting in narrowband but multi-frequency soundscapes (see Figure 2).

The results did not show inter-group differences in the localization accuracy between the Monotonic and the Harmonic encodings. It suggests that adding 2 octaves to the original sound (i.e., the Monotonic encoding) did not modulate the ability to perceive the elevation of the target. Using more complex tones with several sub-octave intervals in the Harmonic encoding might sufficiently modify the sound spectrum to obtain a significant difference with the Monotonic encoding. It could also be interesting to investigate the ability to perceive the elevation of the target with an encoding using sounds containing frequencies higher than the current ceiling frequency (6,000 Hz). However, as mentioned in Section 4.1.1, it seems that the benefits that could arise from the application of the HRTFs on a sound with a broader spectrum could only be perceivable after a long training period.

4.2 Azimuth localization using the visual-to-auditory encodings is accurate but overestimated

Depending on the encoding and the target eccentricity, the magnitude of the azimuth unsigned errors was comprised between 16.78 ± 20.31° and 37.29 ± 20.32°. As a comparison, Mendonça et al. (2013) spatialized white noise sounds using the same HRTFs database and their participants localized the azimuth with a mean unsigned error of 21.3° before the training practice.

In Geronazzo et al. (2018), the azimuth unsigned errors of participants varied between 3.67 ± 2.97° and 35.98 ± 45.32°. In the SSD domain, Scalvini et al. (2022) found a mean azimuth error of 6.72 ± 5.82° in a task consisting in localizing a target with the head. In the current study, after the familiarization, the magnitude of azimuth unsigned errors decreased and was comprised between 12.04 ± 12.05° and 25.06 ± 14.48° depending on the azimuth eccentricity which is comparable to the azimuth unsigned errors found in Geronazzo et al. (2018), without training. In Mendonça et al. (2013), immediately after the training, the mean azimuth unsigned errors also decreased and reached a magnitude of 15.3° which is also comparable to the current results.

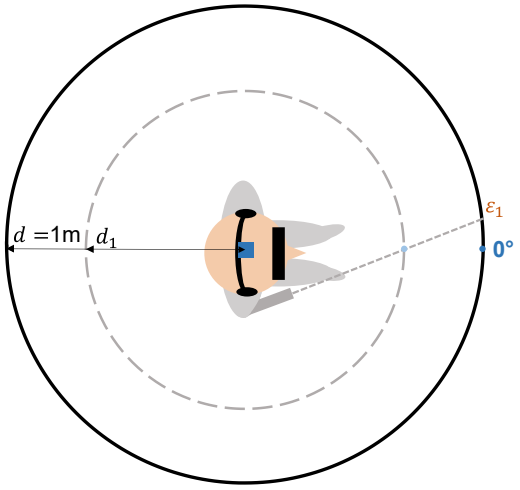
In the current study, without any familiarization, and with the three visual-to-auditory encodings, participants were able to discriminate the different azimuths as suggested by gains higher than the optimal value of 1.0. After the familiarization, and with the three visual-to-auditory encodings, participants were able to localize the azimuth of the target with average azimuth gains comprised between 1.23 and 1.41 which were higher than the null value and than the optimal gain 1.0. It shows that the sound spatialization method used in the current study based on HRTFs from the CIPIC database (Algazi et al., 2001a) partly reproduced the natural cues used in free-field sound azimuth localization. These results are not surprising since azimuth is mainly conveyed through binaural cues including the Interaural Level Difference (ILD) and the Interaural Time Difference (ITD) that reflect audio signal differences between the two ears. ITD is mainly used when the spectral content of the audio signal does not include frequencies higher than 1,500 Hz and ILD is mainly used for frequencies higher than 3,000 Hz (Blauert, 1996). The used frequencies ranged from 250 Hz to about 1,500 Hz with the Monotonic encoding, which is in a frequency domain where ITDs are mainly used to perceive the azimuth. With the Harmonic encoding, that added 2 octaves, the frequency range was between 250 and 6,000 Hz which already contains the ILD frequency domain. The Noise encoding with the broadband sound allows both cues (ITD and ILD) to be fully used, which can theoretically improve azimuth localization accuracy in comparison with sounds with a lower spectral complexity, as previously shown in Morikawa and Hirahara (2013). However, in the current study, these drastic changes in the spectrum did not strongly affect the participants' abilities, and the response patterns were similar. In other words, whatever the spectral complexity of the sound used in the encoding (white noise, complex tones or pure tones), binaural cues could be perceived and interpreted by the participants. It can be noticed that azimuth accuracy seems slightly higher with the two pitch-based encodings (the Harmonic and Monotonic encodings) in comparison with the spatialization-based only encoding (the Noise encoding). We did not find similar results in the scientific literature. This facilitation effect could result from a decrease in the cognitive load when the elevation is conveyed through the pitch modulation. As mentioned above, the pitch-based encodings seem more intuitive to localize the elevation,

therefore it should globally decrease the cognitive load and thus facilitate the processing of the remaining dimension (i.e., the azimuth dimension). This effect does not seem to drastically shape the results and remains to be confirmed by other experiments.

The participants tended to overestimate the lateral position of the lateral targets with the three visual-to-auditory encodings: a shift to the left for targets on the left, and a shift to the right for targets on the right. Some studies also showed an overestimation pattern of lateral sound sources while using non-individualized HRTFs (Wenzel et al., 1993), in a virtual environment while being blindfolded (Ahrens et al., 2019), using ambisonics (Huisman et al., 2021), and even with real sound sources (Oldfield and Parker, 1984; Makous and Middlebrooks, 1990). A possibility to decrease this overestimation might be to rescale the used HRTF positions to fit to the perceived ones. For example one could rescale the azimuth angles of the HRTFs database to compensate for the non-linear shape that was measured as the perceived ones and measure if it could linearize the response profile.

The participants also tended to localize the targets with a leftward bias between -1.2 and -5.03° in average. This systematic error might be due to a wrong auditory localization but also to a misperception of target distance. Geometrical considerations shows that an underestimation of the distance of the sound source would generate a leftward bias as we see in the current results. Since no indication concerning the sound distance was given, the participants could estimate that the sound sources were located closer than one meter. Figure 7 shows the effect of a misperception of target distance on the azimuth localization. However, for the same reason, a distance underestimation would have cause an overestimation of the elevation perception, which we did not measure.

Figure 7. Left shift scheme. If the participant perceived the distance of the target closer than the real target distance (d1 instead of d = 1m), it might induce an increase of the leftward bias (ε1).



4.3 A fast improvement in object localization performance

4.3.1 A short but active familiarization method

After a first practice followed by a very short familiarization, participants' abilities to localize an object with the visual-to-auditory SSD were improved. The elevation gains were improved for all the encodings (especially for pitch-based ones), and for the azimuth, the decrease in the lateral overshoot suggests that the interpretation of acoustic cues provided by the ILD and ITD for the azimuth was improved. Since no feedback was given during the first practice, it can be supposed that the familiarization session mainly contributed to acquire sensorimotor contingencies (Auvray, 2004) through the mean of an audio-motor calibration (Aytekin et al., 2008).

In order to avoid a too long experimental session, we used a short audio-motor familiarization session (60 s) during which participants were active by controlling the position of the target, which is known to improve the positive effect of the training (Aytekin et al., 2008; Hüg et al., 2022). Other familiarization methods have been studied and have shown improvements in the use of SSDs. For example, prior to the experimental task, some studies simultaneously displayed to participants an image and its equivalent soundscape (Ambard et al., 2015; Buchs et al., 2021). In another study (Auvray et al., 2007), participants were enrolled in an intensive training of 3 h. Using only a verbal explanation of the visual-to-auditory encoding scheme as been shown to be efficient to understand the main principles of the encoding scheme (Kim and Zatorre, 2008; Buchs et al., 2021; Scalvini et al., 2022). The aim of the current study was not to directly investigate the effect of a short and active familiarization method on localization performance but it shows that a short practice might be sufficient to acquire the sensorimotor contingencies. The effect of the familiarization remains to be clearly assessed by comparing the efficiency of the existing methods with control conditions in order to optimize the SSD learning.

4.3.2 Calibration of the auditory space improves localization abilities

In the current study, participants were not aware of the size of the VAS neither that the head tracker was associated with a virtual camera capturing and converting into sounds a limited portion of the virtual scene in front of them. They only knew that the virtual target would appear at random locations in their front-field at different azimuth and elevation locations. As a consequence, they also did not know the spatial boundaries of the space where the target could be heard. After a short practice, the participants were able to build an accurate mental spatial representation of the virtual space where the visual-to-auditory encoding took place. For instance, the downward bias in elevation decreased after the familiarization session, suggesting that

participants learned that the VAS was at a higher location. Also the decrease of the overestimation pattern in azimuth suggests that participants learned that the lateral VAS boundaries were closer.

It has to be noticed that the size of the VAS has an influence on the localization accuracy. The biggest the VAS is, the higher the localization error might be. Restricting the field of view of the camera would result in a smaller possible space in which an heard target could be placed, thus resulting in a lower angular error, but as a counterpart, it would cover a smaller subpart of the front-field without moving the head. For instance, for a target placed in a central position, a random pointing in a VAS with a field of view of 45 × 45° (azimuth × elevation) would result in an error in azimuth and elevation with a standard deviation 2 times lower than with a field of view of 90 × 90° while covering a space 4 times smaller. Studying the effect of various VAS sizes in a target localization task in which the user can freely move the head to point to a target as fast as possible would probably give some insights about the optimal VAS size. However, in ecological contexts, a large VAS size would have the advantage of providing auditory information about obstacles placed with a larger eccentricity with respect to the forward direction of the head. For this reason, in a real context of use, this parameter should probably be customizable according to the habit of use.

5. Conclusion

Long trainings are required to master a visual-to-auditory SSD (Kristjánsson et al., 2016) because the used visual-to-auditory encodings are not enough intuitive (Hamilton-Fletcher et al., 2016b). In our study, we investigated several visual-to-auditory encodings in order to develop a SSD whose auditory information could quickly be interpreted to localize obstacles. In line with previous studies, our results suggest that a visual-to-auditory SSD based on the creation of a VAS is efficient to convey visuo-spatial information about azimuth through soundscapes. Our study shows that a pitch-based elevation mapping can be easily learn to compensate for elevation localization impairments due to the use of non-individualized HRTFs in the creation process of the VAS. Despite a very short period of practice, the participants were able to improve their interpretation of the used acoustic cues both for the azimuth and the elevation encoding schemes.

Ethics statement

The studies involving human participants were reviewed and approved by Comité d'Ethique pour la Recherche de Université Bourgogne Franche-Comté. The patients/participants provided their written informed consent to participate in this study.

Author contributions

CB and MA contributed to conception and design of the study and interpreted the data. CB executed the study and was responsible for data analysis and wrote the first draft of the manuscript in closed collaboration with MA. FS, CM, and JD provided important feedback. All authors have read, approved the manuscript, and contributed substantially to it.

Funding

This research was funded by the Conseil Régional de Bourgogne Franche-Comté (2020_0335), France and the Fond Européen de Développement Régional (FEDER) (BG0027904).

Acknowledgments

Thanks to the Conseil Régional de Bourgogne Franche-Comté, France and the Fond Européen de Développement Régional (FEDER) for their financial support. We thank the Université de Bourgogne and le Centre National de la Recherche Scientifique (CNRS) for the providing administrative support and the infrastructure.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

6. References

Abboud, S., Hanassy, S., Levy-Tzedek, S., Maidenbaum, S., & Amedi, A. (2014). EyeMusic : introducing a "visual" colorful experience for the blind using auditory sensory substitution. *Restorative Neurology and Neuroscience, 32*(2), 247–257. https://doi.org/10.3233/rnn-130338

Ahrens, A., Lund, K. D., Marschall, M., & Dau, T. (2019). Sound source localization with varying amount of visual information in virtual reality. *PLOS ONE, 14*(3), e0214603. https://doi.org/10.1371/journal.pone.0214603

Algazi, V. R, Duda, R., Thompson, D., and Avendano, C. (2001a). "The CIPIC HRTF database," *in Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics* (New Platz, NY:L IEEE), 99–102.

Algazi, V. R., Avendaño, C., & Duda, R. O. (2001). Elevation localization and head-related transfer function analysis at low frequencies. Journal of the Acoustical Society of America, 109(3), 1110–1122. https://doi.org/10.1121/1.1349185

Ambard, M., Benezeth, Y., and Pfister, P. (2015). Mobile video-to-audio transducer and motion detection for sensory substitution. *Frontiers in ICT 2*, 20. https://doi.org/10.3389/fict.2015.00020

Asano, F., Suzuki, Y., & Sone, T. (1990). Role of spectral cues in median plane localization. *Journal of the Acoustical Society of America, 88*(1), 159–168. https://doi.org/10.1121/1.399963

Auvray, M. (2004). *Immersion et perception spatiale. L'exemple des dispositifs de substitution sensorielle* (Ph.D. thesis). Ecole des Hautes Etudes en Sciences Sociales, Paris.

Auvray, M., Hanneton, S., & O'Regan, J. K. (2007). Learning to perceive with a visuo-auditory substitution system : localisation and object recognition with 'The Voice'. *Perception, 36*(3), 416–430. https://doi.org/10.1068/p5631

Aytekin, M., Moss, C. F., & Simon, J. Z. (2008). A sensorimotor approach to sound localization. *Neural Computation, 20*(3), 603–635. https://doi.org/10.1162/neco.2007.12-05-094

Bauer, R. W., Matuzsa, J. L., Blackmer, R. F., and Glucksberg, S. (1966). Noise localization after unilateral attenuation. *Journal of the Acoustical Society of America, 40*, 441–444. https://doi.org/10.1121/1.1910093

Best, V., Baumgartner, R., Lavandier, M., Majdak, P., and Kopčo, N. (2020). Sound externalization: a review of recent research. *Trends in hearing*, 24, 233121652094839. https://doi.org/10.1177/2331216520948390

Blauert, J. (1996). Spatial hearing. *In The MIT Press eBooks*. https://doi.org/10.7551/mitpress/6391.001.0001

Brown, D., Macpherson, T., and Ward, J. (2011). Seeing with sound? Exploring different characteristics of a visual-to-auditory sensory substitution device. *Perception 40,* 1120–1135. https://doi.org/10.1068/p6952

Buchs, G., Haimler, B., Kerem, M., Maidenbaum, S., Braun, L., and Amedi, A. (2021). A self-training program for sensory substitution devices. *PLOS ONE 16*, e0250281. https://doi.org/10.1371/journal.pone.0250281

Caraiman, S., Morar, A., Owczarek,M., Burlacu, A., Rzeszotarski, D., Botezatu,N., et al. (2017). "Computer vision for the visually impaired: the sound of vision system," *in 2017 IEEE International Conference on Computer Vision Workshops (ICCVW)* (Venice: IEEE), 1480–1489.

Commère, L., Wood, S. U. N., & Rouat, J. (2020). Evaluation of a vision-to-audition substitution system that provides 2D WHERE information and fast user learning. ArXiv:2010.09041. http://arxiv.org/abs/2010.09041

Deroy, O., Fasiello, I., Hayward, V., and Auvray,M. (2016). Differentiated audio-tactile correspondences in sighted and blind individuals. *Journal of Experimental Psychology : Human Perception and Performance, 42*(8), 1204–1214. https://doi.org/10.1037/xhp0000152

Deroy, O., Fernandez-Prieto, I., Navarra, J., and Spence, C. (2018). "Unraveling the paradox of spatial pitch," *in Spatial Biases in Perception and Cognition*, 1st Edn, ed T. L. Hubbard (New York, NY: Cambridge University Press), 77–93.

Evans, K. K., and Treisman, A. (2011). Natural cross-modal mappings between visual and auditory features. *Journal of Vision. 10*(1), 6. https://doi.org/10.1167/10.1.6

Gardner, M. B. (1973). Some monaural and binaural facets of median plane localization. *Journal of the Acoustical Society of America, 54*, 1489–1495. https://doi.org/10.1121/1.1914447

Geronazzo, M., Sikstrom, E., Kleimola, J., Avanzini, F., de Gotzen, A., and Serafin, S. (2018). "The impact of an accurate vertical localization with HRTFs on short explorations of immersive virtual reality scenarios," *in 2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)* (Munich: IEEE), 90–97.

Hamilton-Fletcher, G., Mengucci, M., and Medeiros, F. (2016a). Synaestheatre: sonification of coloured objects in space. *Brighton: International Conference on Live Interfaces.*

Hamilton-Fletcher, G., Obrist, M., Watten, P., Mengucci, M., and Ward, J. (2016b). “"I always wanted to see the night sky": blind user preferences for sensory substitution devices,” in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, CA: ACM), 2162–2174.

Hanneton, S., Auvray, M., and Durette, B. (2010). The Vibe: a versatile vision-to-audition sensory substitution device. *Applied Bionics and Biomechanics*, *7*(4), 269–276. https://doi.org/10.1155/2010/282341

Hebrank, J., and Wright, D. (1974). Spectral cues used in the localization of sound sources on the median plane. *Journal of the Acoustical Society of America, 56*(6), 1829–1834. https://doi.org/10.1121/1.1903520

Howard, D. M., and Angus, J. (2009). *Acoustics ans Psychoacoustics,* 4th Edn. Oxford: Focal Press.

Hüg, M. X., Bermejo, F., Tommasini, F. C., and Di Paolo, E. A. (2022). Effects of guided exploration on reaching measures of auditory peripersonal space. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.983189

Huisman, T., Ahrens, A., and MacDonald, E. (2021). Ambisonics sound source localization with varying amount of visual information in virtual reality. *Frontiers in Virtual Reality 2*. https://doi.org/10.3389/frvir.2021.722321

Jicol, C., Lloyd-Esenkaya, T., Proulx, M. J., Lange-Smith, S., Scheller, M., O’Neill, E., et al. (2020). Efficiency of sensory substitution devices alone and in combination with self-motion for spatial navigation in sighted and visually impaired. *Frontiers in Psychology, 11*. https://doi.org/10.3389/fpsyg.2020.01443

Kim, J.-K., and Zatorre, R. J. (2008). Generalized learning of visual-to- auditory substitution in sighted individuals. *Brain Research, 1242*, 263–275. https://doi.org/10.1016/j.brainres.2008.06.038

Kristjánsson, Á., Moldoveanu, A., Jóhannesson, Ó. I., Bălan, O., Spagnol, S., Valgeirsdóttir, V. V., & Unnþórsson, R. (2016). Designing sensory-substitution devices: Principles, pitfalls and potential1. *Restorative Neurology and Neuroscience, 34*, 769–787. https://doi.org/10.3233/RNN-160647

Kumar, S., Forster, H. M., Bailey, P., and Griffiths, T. D. (2008). Mapping unpleasantness of sounds to their auditory representation. *Journal of the Acoustical Society of America, 124*, 3810–3817. https://doi.org/10.1121/1.3006380

Kumpik, D. P., Kacelnik, O., and King, A. J. (2010). Adaptive reweighting of auditory localization cues in response to chronic unilateral earplugging in humans. *The Journal of Neuroscience, 30*, 4883–4894. https://doi.org/10.1523/JNEUROSCI.5488-09.2010

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). LMERTest Package : Tests in Linear Mixed Effects models. *Journal of Statistical Software, 82*(13). https://doi.org/10.18637/jss.v082.i13

Lenth, R. V. (2022). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.7.4–1.

Levy-Tzedek, S., Hanassy, S., Abboud, S., Maidenbaum, S., and Amedi, A. (2012). Fast, accurate reaching movements with a visual-to-auditory sensory substitution device. *Restorative Neurology and Neuroscience, 30*, 313–323. https://doi.org/10.3233/RNN-2012-110219

Maidenbaum, S., Abboud, S., and Amedi, A. (2014). Sensory substitution: closing the gap between basic research and widespread practical visual rehabilitation. *Neuroscience & Biobehavioral Reviews, 41,* 3–15. https://doi.org/10.1016/j.neubiorev.2013.11.007

Maidenbaum, S., and Amedi, A. (2019). "Standardizing visual rehabilitation using simple virtual tests," *in 2019 International Conference on Virtual Rehabilitation (ICVR)* (Tel Aviv: IEEE), 1–8.

Majdak, P., Walder, T., and Laback, B. (2013). Effect of long-term training on sound localization performance with spectrally warped and band-limited head-related transfer functions. *Journal of the Acoustical Society of America, 134*, 2148–2159. https://doi.org/10.1121/1.4816543

Makous, J. C., and Middlebrooks, J. C. (1990). Two-dimensional sound localization by human listeners. *Journal of the Acoustical Society of America, 87*, 2188–2200. https://doi.org/10.1121/1.399186

Meijer, P. (1992). An experimental system for auditory image representations. *IEEE Transactions on Biomedical Engineering, 39*(2), 112–121. https://doi.org/10.1109/10.121642

Mendonça, C., Campos, G., Dias, P., and Santos, J. A. (2013). Learning auditory space: generalization and long-term effects. *PLOS ONE* 8, e77900. https://doi.org/10.1371/journal.pone.0077900

Mhaish, A., Gholamalizadeh, T., Ince, G., and Duff, D. J. (2016). "Assessment of a visual to spatial-audio sensory substitution system*," in 2016 24th Signal Processing and Communication Application Conference (SIU)* (Zonguldak: IEEE), 245–248.

Middlebrooks, J. C. (1999). Individual differences in external-ear transfer functions reduced by scaling in frequency. *Journal of the Acoustical Society of America, 106*, 1480–1492. https://doi.org/10.1121/1.427176

Middlebrooks, J. C., and Green, D. M. (1990). Directional dependence of interaural envelope delays. *Journal of the Acoustical Society of America, 87*, 2149–2162. https://doi.org/10.1121/1.399183

Middlebrooks, J. C., & Green, D. M. (1991). Sound localization by human listeners. *Annual Review of Psychology, 42*(1), 135–159. https://doi.org/10.1146/annurev.ps.42.020191.001031

Miller, J. (1991). Channel interaction and the redundant-targets effect in bimodal divided attention. *Journal of Experimental Psychology: Human Perception and Performance, 17*(1), 160–169. https://doi.org/10.1037/0096-1523.17.1.160

Morikawa, D., & Hirahara, T. (2013). Effect of head rotation on horizontal and median sound localization of band-limited noise. *Acoustical Science and Technology, 34*(1), 56–58. https://doi.org/10.1250/ast.34.56

Oldfield, S. R., and Parker, S. P. A. (1984). Acuity of sound localisation: a topography of auditory space. I. Normal hearing conditions. *Perception 13*, 581–600. https://doi.org/10.1068/p130581

Pourghaemi, H., Gholamalizadeh, T., Mhaish, A., Duff, D. J., and Ince, G. (2018). Realtime shape-based sensory substitution for object localization and recognition. *Proceedings of the 11th International Conference on Advances in Computer-Human Interactions.*

Proulx, M. J., Stoerig, P., Ludowig, E., and Knoll, I. (2008). Seeing 'where through the ears: effects of learning-by-doing and long-term sensory deprivation on localization based on image-to-sound substitution. *PLOS ONE, 3*(3), e1840. https://doi.org/10.1371/journal.pone.0001840

Real, S., and Araujo, A. (2021). VES: a mixed-reality development platform of navigation systems for blind and visually impaired. *Sensors 21*, 6275. https://doi.org/10.3390/s21186275

Richardson, M., Thar, J., Alvarez, J., Borchers, J., Ward, J., and Hamilton-Fletcher, G. (2019). How much spatial information is lost in the sensory substitution process? Comparing visual, tactile, and auditory approaches. *Perception 48*, 1079–1103. https://doi.org/10.1177/0301006619873194

Romigh, G. D., Simpson, B., and Wang, M. (2017). Specificity of adaptation to non-individualized head-related transfer functions. *Journal of the Acoustical Society of America, 141*, 3974–3974. https://doi.org/10.1121/1.4989065

Rouat, J., Lescal, D., and Wood, S. (2014). Handheld Device for substitution from vision to audition. *New York, NY: 20th International Conference on Auditory Display*.

Rusconi, E., Kwan, B., Giordano, B., Umilta, C., and Butterworth, B. (2006). Spatial representation of pitch height: the SMARC effect. *Cognition 99*, 113–129. https://doi.org/10.1016/j.cognition.2005.01.004

Scalvini, F., Bordeau, C., Ambard, M., Migniot, C., and Dubois, J. (2022). "Low-latency human-computer auditory interface based on real-time vision analysis," *in ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Singapore: IEEE), 36–40.

Shinn-Cunningham, B. G., Durlach, N. I., and Held, R. M. (1998). Adapting to supernormal auditory localization cues. I. Bias and resolution. *Journal of the Acoustical Society of America, 103,* 3656–3666. https://doi.org/10.1121/1.423088

Sodnik, J., Sušnik, R., Štular, M., & Tomažič, S. (2005). Spatial sound resolution of an interpolated HRIR library. *Applied Acoustics, 66*(1), 1219–1234. https://doi.org/10.1016/j.apacoust.2005.04.003

Spence, C. (2011). Crossmodal Correspondences: A tutorial review. *Attention, perception & psychophysics, 73*(4), 971–995. https://doi.org/10.3758/s13414-010-0073-7

Spence, C., and Deroy, O. (2013). How automatic are crossmodal correspondences? *Consciousness and Cognition, 22*(1), 245–260. https://doi.org/10.1016/j.concog.2012.12.006

Steinmetz, C. J., and Reiss, J. D. (2021). "Pyloudnorm: a simple yet flexible loudness meter in python," *in 150th AES Convention*. Available online at: https://csteinmetz1.github.io/pyloudnorm-eval/

Stevens, S. S., Volkmann, J., and Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America,* 8, 185–190. https://doi.org/10.1121/1.1915893

Stiles, N. R. B., and Shimojo, S. (2015). Auditory sensory substitution is intuitive and automatic with texture stimuli. *Scientific Report, 5*. https://doi.org/10.1038/srep15628

Team, R. C. (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R Core Team.

Voss, P. (2016). Auditory spatial perception without vision. *Frontiers in Psychology, 07*. https://doi.org/10.3389/fpsyg.2016.01960

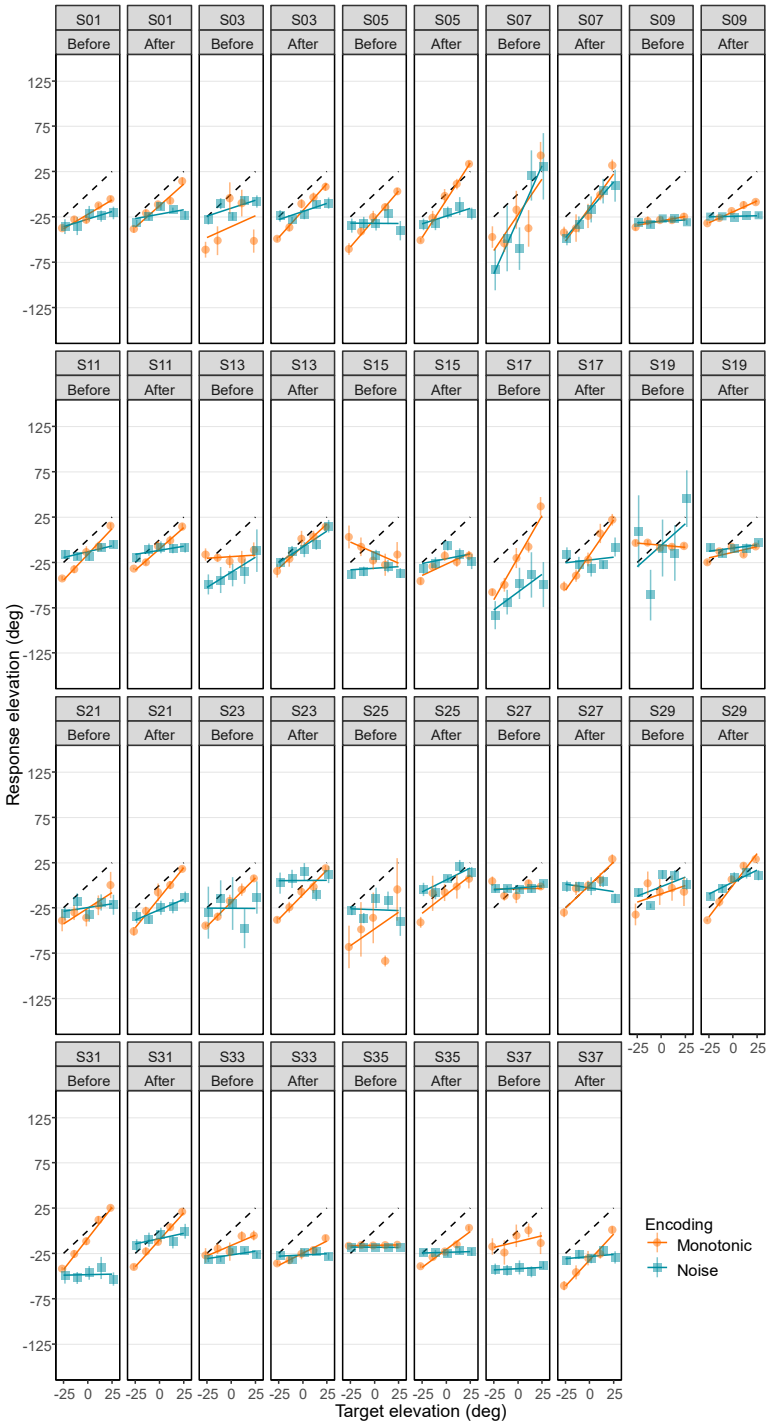
Wenzel, E. M., Arruda, M., Kistler, D. J., and Wightman, F. L. (1993). Localization using nonindividualized head-related transfer functions. *Journal of the Acoustical Society of America, 94*, 111–123. https://doi.org/10.1121/1.407089

Xu, S., Li, Z., and Salvendy, G. (2007). "Individualization of head-related transfer function for three-dimensional virtual auditory display: a review," *in Proceedings of the 2nd International Conference on Virtual Reality, ICVR'07* (Berlin; Heidelberg: Springer-Verlag), 397–407.

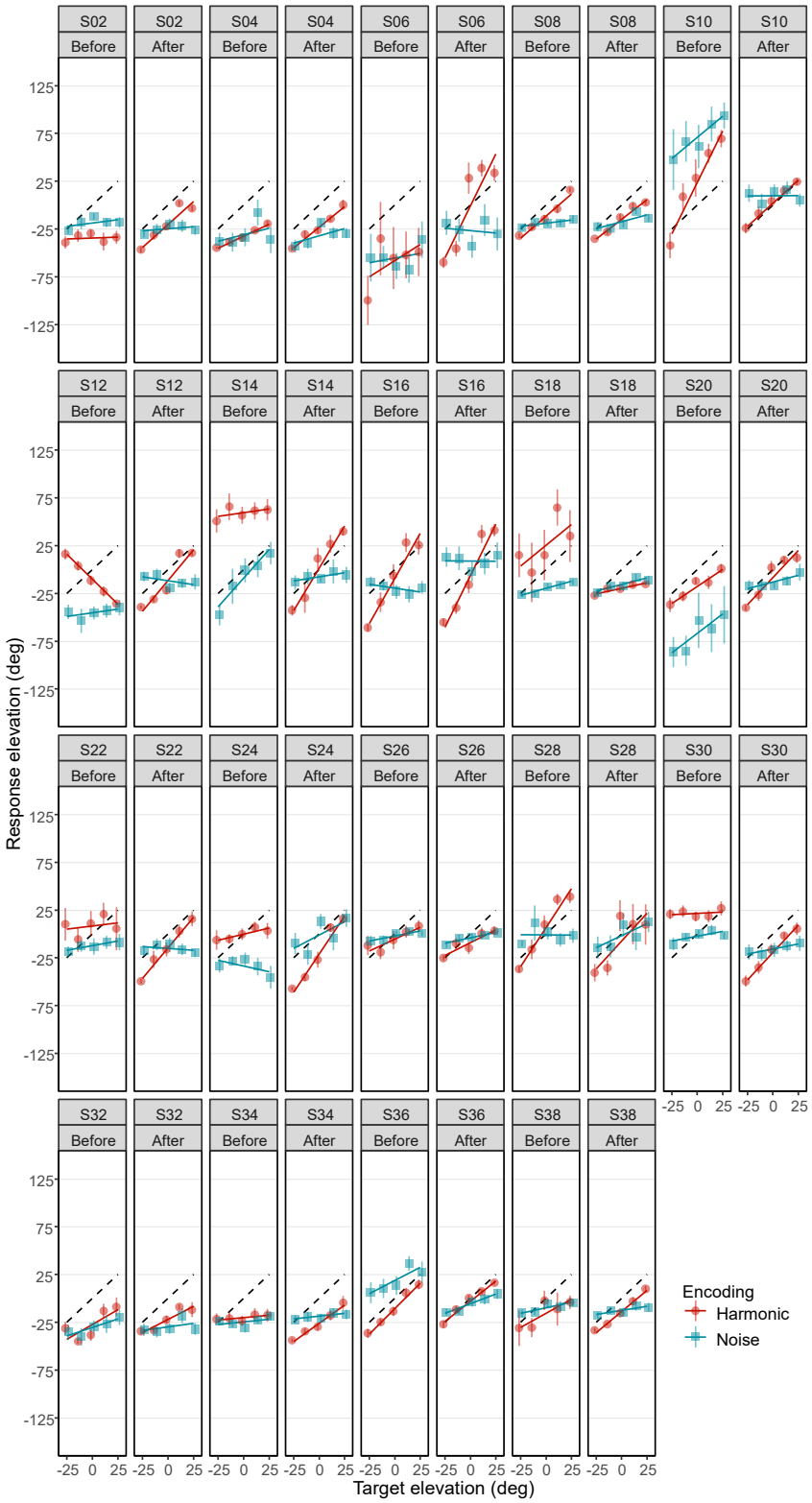
Zwicker, E. (1961). Subdivision of the audible frequency range into critical bands (Frequenzgruppen). *Journal of the Acoustical Society of America*, *33*, 248–248. https://doi.org/10.1121/1.1908630

7. Supplementary material

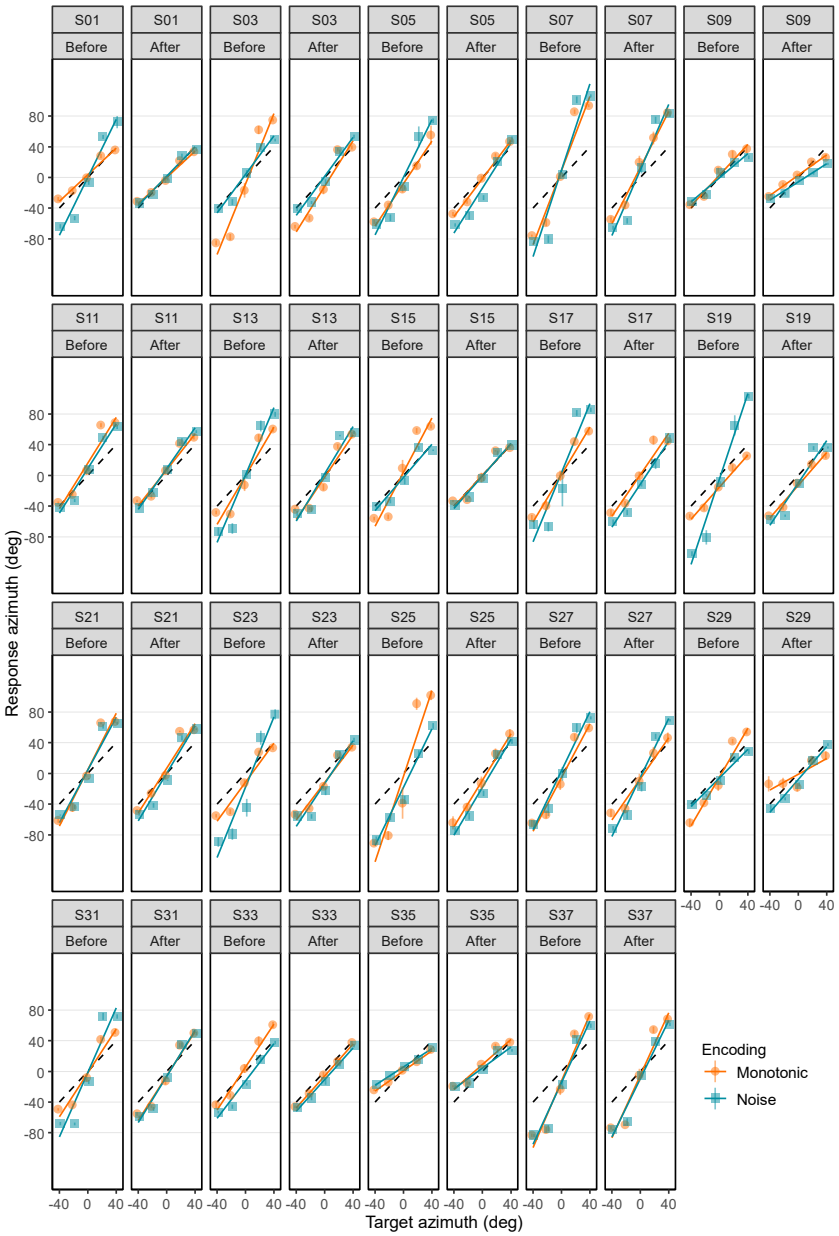
Supplementary Figure S1. Elevation response position as a function of target elevation for each participant of the Monotonic group. Mean elevation response positions (in degree) before (left) and after (right) are represented separately for the Noise (blue squares) and the Monotonic (orange circles) encodings. Error bars shows standard error of elevation response position. Solid lines represent the elevation gains with the Noise (blue) and Monotonic (orange) encodings. Black dashed lines indicate the optimal elevation gain 1.0.



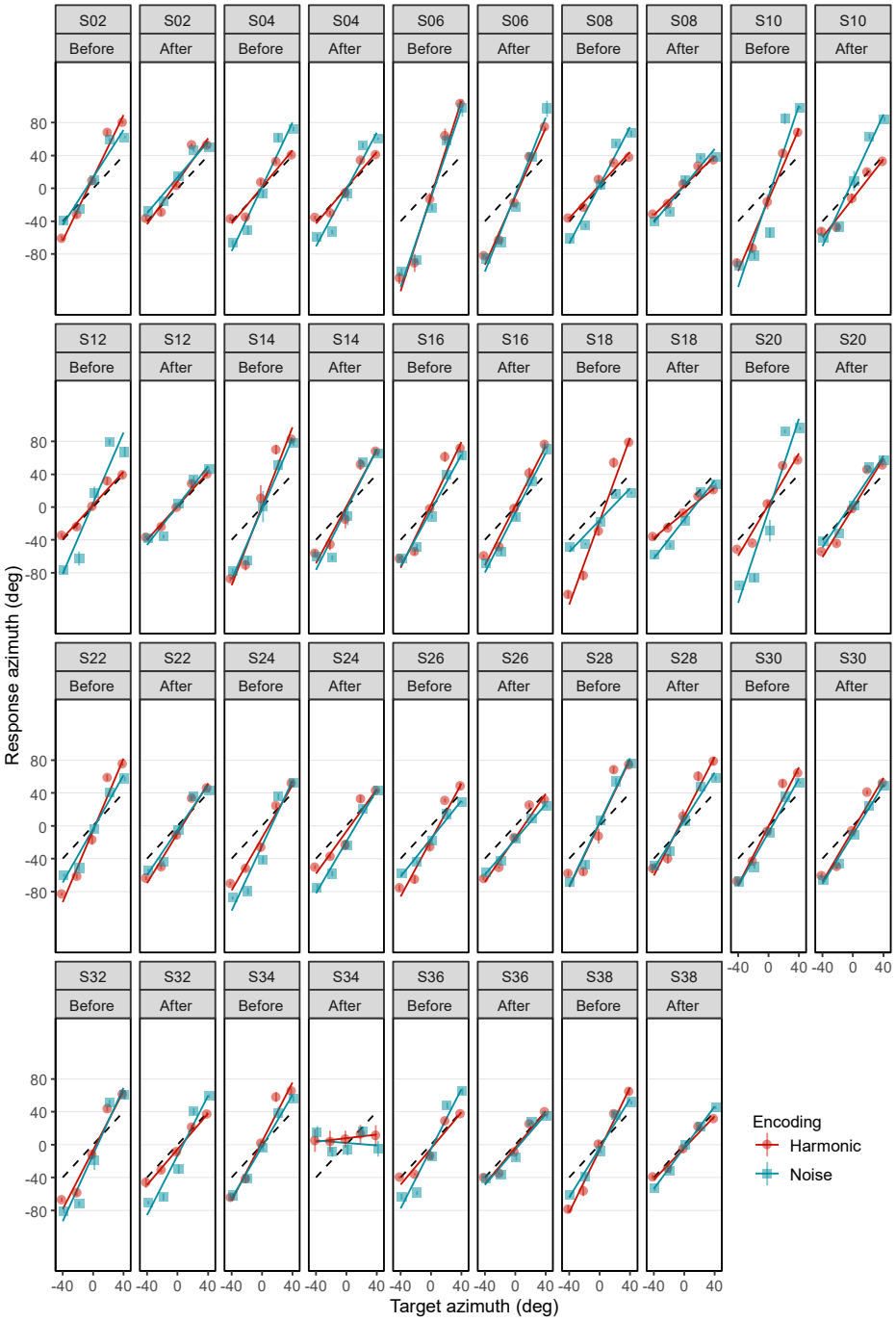
Supplementary Figure S2. Elevation response position as a function of target elevation for each participant of the Harmonic group. Mean elevation response positions (in degree) before (left) and after (right) are represented separately for the Noise (blue squares) and the Harmonic (red circles) encodings. Error bars shows standard error of elevation response position. Solid lines represent the elevation gains with the Noise (blue) and Harmonic (red) encodings. Black dashed lines indicate the optimal elevation gain 1.0.



Supplementary Figure S3. Azimuth response position as a function of target azimuth for each participant of the Monotonic group. Mean azimuth response positions (in degree) before (left) and after (right) are represented separately for the Noise (blue squares) and the Monotonic (orange circles) encodings. Error bars shows standard error of azimuth response position. Solid lines represent the azimuth gains with the Noise (blue) and Monotonic (orange) encodings. Black dashed lines indicate the optimal azimuth gain 1.0.



Supplementary Figure S4. Azimuth response position as a function of target azimuth for each participant of the Harmonic group. Mean azimuth response positions (in degree) before (left) and after (right) are represented separately for the Noise (blue squares) and the Harmonic (red circles) encodings. Error bars shows standard error of azimuth response position. Solid lines represent the azimuth gains with the Noise (blue) and Harmonic (red) encodings. Black dashed lines indicate the optimal azimuth gain 1.0.



1.3. Synthèse

L'Étude 1 avait pour objectif d'évaluer les capacités de localisation d'objets avec le DSS dans les dimensions de l'azimut et de l'élévation en comparant trois schémas d'encodage reposant sur la spatialisation en azimut et en élévation d'une source sonore avec des HRTFs non-individualisées. Un premier schéma d'encodage reposait uniquement sur des indices acoustiques spatiaux spatialisant en azimut et en élévation une source sonore à large spectre (schéma d'encodage *Noise*). Les deux autres schémas d'encodage intégraient la modulation d'un indice acoustique supplémentaire souvent utilisé dans les DSSs existants et qui est impliqué dans une correspondance cross-modale audio-visuelle : la hauteur tonale. Ces deux schémas d'encodage reposaient donc sur la spatialisation en azimut et en élévation de sources sonores avec un spectre étroit (des tonalités dont la hauteur tonale augmentait avec l'augmentation de l'élévation), avec soit des tonalités pures pour le schéma d'encodage *Monotonic*, ou des tonalités complexes pour le schéma d'encodage *Harmonic*.

Lors d'une tâche de pointage, après une courte familiarisation audio-motrice, les résultats ont premièrement montré que les capacités de localisation de l'élévation avec un DSS utilisant un schéma d'encodage reposant sur la simulation de sources sonores avec des HRTFs non-individualisées sont altérées. Ces résultats confirment que l'utilisation d'indices acoustiques spatiaux dans le schéma d'encodage d'un DSS fait face à des limites perceptives similaires aux limites de perception spatiale auditive.

D'une autre part, les résultats mettent en évidence de meilleures capacités de localisation de l'élévation d'un objet avec le schéma d'encodage modulant la hauteur tonale. Ces meilleures performances de localisation peuvent être attribuées à un effet facilitateur de la correspondance cross-modale entre hauteur tonale et élévation visuelle. En revanche, la complexité spectrale des tonalités ne modulait pas les performances de localisation. Ces résultats peuvent être expliqués soit par l'utilisation de tonalités à spectre trop étroit ou pas assez complexes dans le schéma d'encodage *Harmonic*, soit par un effet plafond avec des performances maximales déjà atteintes avec le schéma d'encodage *Monotonic*. En ce qui concerne les capacités de localisation de l'azimut, les résultats confirment que les indices acoustiques spatiaux sont efficaces pour transmettre des informations spatiales sur l'azimut d'un objet à travers le paysage sonore d'un DSS.

Dans le contexte du développement de DSS vision-vers-audition pour l'aide à la locomotion et à la localisation d'obstacles, il est important d'estimer correctement à la fois la position latérale des obstacles mais aussi leur hauteur pour réduire la probabilité de collision avec les obstacles. Les résultats de l'Étude 1 suggèrent qu'aux premiers stades de l'utilisation du DSS, l'utilisation de la hauteur tonale comme indice acoustique dans les DSSs permet de compenser les

limites perceptives de l'élévation lorsqu'une spatialisation avec des HRTFs non-individualisées est employée, et que l'utilisation d'indices spatiaux pour la dimension de l'azimut est efficace mais qu'il serait intéressant de tenter de réduire la surestimation de la latéralité des obstacles qui pourrait mener à des collisions.

2. Étude 2

Perception de la distance avec un dispositif de substitution sensorielle vision-vers-audition : compenser le biais de compression avec l'enveloppe sonore

Distance perception with a visual-to-auditory substitution device: compensating for the compression bias using sound envelope

Camille Bordeau, Florian Scalvini, Cyrille Migniot, Julien Dubois and Maxime Ambard

À soumettre

2.1. Résumé

Éviter les obstacles lors des déplacements pédestres reste un défi central pour les personnes non-voyantes. Les dispositifs de substitution sensorielle vision-vers-audition utilisent différents indices acoustiques pour transmettre des informations relatives à la distance des objets composant la scène environnante. Généralement, ils reproduisent des indices acoustiques écologiques (par exemple, l'intensité, les réverbérations) ou synthétisent des indices plus arbitraires (par exemple, le taux de répétition des bips, la hauteur des sons). La présente étude vise à comparer les performances de perception de la distance d'objets avec un dispositif de substitution utilisant deux schémas d'encodage : l'un utilisant la modulation de l'intensité sonore, et l'autre couplant la modulation de l'intensité sonore avec et la modulation de l'amplitude de l'enveloppe. Les performances de localisation ont été évaluées avec des participants aux yeux bandés dans un environnement virtuel, à la fois avec une tâche de localisation basée sur une nouvelle méthode de pointage, et avec une tâche de discrimination de la distance.

Les résultats mettent en évidence un biais de compression de la distance avec les deux schémas d'encodage, par ailleurs réduit avec le schéma d'encodage utilisant la modulation de l'amplitude de l'enveloppe, ce qui suggère une amélioration de la performance de localisation. Les capacités des participants à évaluer la distance relative de deux cibles étaient également meilleures avec le schéma d'encodage utilisant la modulation de l'amplitude de l'enveloppe.

Ce travail décrit un nouveau protocole (la méthode du pointage au sol) pour l'évaluation des capacités de perception de la distance dans le contexte de la conception de la substitution sensorielle vision-vers-audition. Il démontre l'efficacité des deux schémas d'encodage utilisés pour transmettre les informations sur la distance et montre que la modulation de l'enveloppe sonore pourrait être utilisée pour réduire le biais de compression généralement observé dans la perception des distances. Cela pourrait contribuer à réduire la probabilité de collision des personnes non-voyantes avec des obstacles lors de leurs déplacements pédestres.

2.2. Article

Distance perception with a visual-to-auditory substitution device: compensating for the compression bias using sound envelope

Camille Bordeau[1], Florian Scalvini[2], Cyrille Migniot[2], Julien Dubois[2] and Maxime Ambard[1]

[1] LEAD-CNRS UMR5022, Université de Bourgogne, Dijon, France
[2] ImViA EA 7535, Université de Bourgogne, Dijon, France

Abstract

Introduction: Avoiding obstacles while walking is a challenge for blind people. Visual-to-auditory sensory substitution devices are assistive tools designed to help the blind perceive the surrounding environment by converting visual features into soundscapes. Such devices use a camera to acquire spatial information about the position of the surrounding objects in the three-dimensional space and convey this information using auditory features. To convey distance information, they usually either mimic ecological acoustic cues (e.g., intensity, reverberations) or synthesize more arbitrary ones (e.g., beep repetition rate, pitch).

Method: The aim of the current study was to compare distance perception performance with a substitution device using two encoding schemes: one using intensity modulation and the other using intensity and envelope amplitude modulation. The performance of blindfolded participants in a virtual environment was assessed both with a localization task based on a new pointing method and with a distance discrimination task.

Results: A distance compression bias was observed with both encoding schemes. However, it was reduced with encoding scheme using envelope amplitude modulation, suggesting that localization performance was improved. Participants also judged the relative distance of two targets better with the encoding using envelope amplitude modulation.

Discussion: This work describes a new protocol (the floor pointing method) for assessing distance perception abilities in the context of visual-to-auditory sensory substitution design. It demonstrates the effectiveness of the two encoding schemes used to convey distance information and shows that sound envelope modulation could be used to reduce the compression bias that is usually observed in distance perception. This could help reduce the likelihood of colliding with obstacles when walking in the street.

Keywords: Sensory substitution, image-to-sound conversion, absolute distance perception, visual impairment, sound intensity, sound envelope, localization, distance discrimination

1. Introduction

Visual-to-auditory sensory substitution devices (SSDs) are assistive tools developed for the blind that convert visual information into soundscapes in order to transmit spatial information about the surrounding environment. They are intended to assist the blind in daily tasks such as object recognition, localization, and obstacle avoidance. Although the study by Renier & De Volder (2010) has shown that it is, to some extent, possible to perceive the distance of an object through soundscapes that do not explicitly transmit depth information, the accurate perception of this information is very important to minimize the occurrence of collisions. Many visual-to-auditory SSDs therefore explicitly convey distance information through dedicated synthesized acoustic cues.

Sound intensity modulation is often used to convey information about distance by increasing sound intensity as the distance decreases (Hamilton-Fletcher et al., 2022; Mhaish et al., 2016; Neugebauer et al., 2020; Ribeiro et al., 2012; Stoll et al., 2015). Intensity modulation has the advantage of simulating physical effects that are naturally used by humans during auditory localization (Zahorik, 2005) as well as of being relatively easy to synthesize (Shinn-Cunningham, 2000). To improve the accuracy of this simulation, some SSDs combine intensity modulation with the use of reverberation cues (Hamilton-Fletcher et al., 2022; Ribeiro et al., 2012). However, studies have shown that users of SSDs that exploit only these acoustic features may find it difficult to perceive distance (Commere & Rouat, 2023; Neugebauer et al., 2020). For example, some participants in Neugebauer et al. (2020) reported having difficulties perceiving the distance of obstacles with an SSD that conveyed distance by means of sound intensity modulation. In another study, Commere & Rouat (2023) reported lower performance in a distance localization task when distance was conveyed using intensity modulation compared to modulation by means of pitch or beep repetition. Moreover, the perception of the distance of real or virtual sound sources is often compressed, with near distances being overestimated and far distances being underestimated (Kolarik et al., 2016; Zahorik, 2005). In the context of SSD, this would result in the overestimation of the distance to nearby objects and increase the probability of collision.

In contrast, other studies on sensory substitution or sensory augmentation have shown the efficiency of other auditory features that are not usually used to perceive the distance of a real sound source, such as beep repetition rate (Commere & Rouat, 2023; Kayukawa et al., 2019; Parseihian et al., 2012), delay (Negen et al., 2018, 2023), pitch modulation (Aladren et al., 2016; Commere & Rouat, 2023), or signal-to-noise ratio (Commere & Rouat, 2023). Although efficient, these artificial sound effects can hardly be incorporated in the sonification scheme of depth maps, in which each pixel conveys three-dimensional information. The sonification of such depth maps is mostly undertaken at a relatively high framerate (about 30 Hz) and uses a combination of

spatialization with HRTFs for the azimuth dimension and pitch modulation for the elevation dimension (Bordeau et al., 2023; Hamilton-Fletcher, Obrist, et al., 2016; Hamilton-Fletcher et al., 2022; Stoll et al., 2015). It is no easy task to find a sound effect that improves the distance perception of depth maps without excessively disrupting the encoding scheme used for azimuth and elevation. In this study, we propose sound envelope modulation as a promising acoustic feature for conveying distance information in the context of SSD.

The sound envelope describes the temporal distribution of energy between the onset (attack time) and the offset (release time) of the acoustic signal (Begault, 1995). To a certain extent, envelope amplitude modulation preserves pitch perception (Rossing & Houtsma, 1986) by relying on the temporal fine structure and only influences the perception of timbre (Plack & Oxenham, 2006). For instance, sounds with an abrupt damped envelope are perceived with a percussive timbre (Schutz & Gillard, 2020). Envelope amplitude modulation therefore seems to be a good candidate for use in an SSD since it should not overly disrupt the perception of pitch, which is often used in SSDs as an acoustic cue to convey information about other spatial dimensions. Pitch modulation is indeed used in many SSDs to convey the elevation dimension (Abboud et al., 2014; Ambard et al., 2015; Capelle et al., 1998; Cronly-Dillon et al., 1999; Hamilton-Fletcher et al., 2022; Hamilton-Fletcher, Mengucci, et al., 2016; Hanneton et al., 2010; Meijer, 1992; Neugebauer et al., 2020; Stoll et al., 2015) or more rarely the azimuth dimension (Capelle et al., 1998). Moreover, envelope modulation preserves sound source localization capabilities for broadband sounds (Yost, 2017) and has even been shown to improve the accuracy of tone localization in the same study. In the context of SSD, envelope amplitude modulation therefore seems to be compatible with an encoding scheme based on tones that are spatialized with HRTFs.

Sound envelope amplitude modulation has the additional advantage of potentially eliciting spatial audiovisual effects. It can, for example, result in the perception of visual spatial events, as in the audiovisual bounce-inducing effect, which refers to the tendency to perceive two identical disks as bouncing instead of crossing when a sound is presented at the time when they cross (Sekuler et al., 1997). It has further been shown that a sound with a damped envelope tends to increase the audiovisual bounce-inducing effect compared with a ramped envelope (Grassi & Casco, 2009).

However, envelope modulation is known to influence other perceptual properties of a sound, such as duration and loudness and, in some configurations, even pitch. It has, for example, been shown that percussive tones tend to be perceived as being shorter than flat tones (Vallet et al., 2014) and ramped broadband sounds tend to be perceived as being longer than damped broadband sounds (Ries et al., 2008). Loudness can also be affected, as suggested by ramped

sounds, which tend to be perceived as being louder than damped sounds when tones (Neuhoff, 1998) or broadband sounds are used (Ries et al., 2008). Rossing & Houtsma (1986) reported that the pitch of short tones with an exponential modulated envelope can be perceived as being higher than that of flat tones with the same frequency. However, these frequency shifts seem to be linked to changes in the average intensity of the acoustic signal, and pitch discrimination abilities are preserved. These results show that envelope modulation influences the perception of many sound attributes, while nevertheless preserving central auditory features that are used in the encoding schemes of most SSDs (pitch and spatial cues) and have the potential to elicit spatial effects. Its efficiency in encoding the distance of objects in the context of an SSD therefore also needs to be assessed.

Various tasks have been used to assess the ability to perceive the distance of an object using a visual-to-auditory SSD. These have included positioning tasks (Commere & Rouat, 2023), direct pointing methods (Auvray et al., 2007; Renier & De Volder, 2010), identification tasks (Mhaish et al., 2016) or an analogical scale (Bazilinskyy et al., 2016) for absolute distance perception, and distance discrimination tasks (Commere & Rouat, 2023; Richardson et al., 2019) for relative distance perception. Similar methods have been used to evaluate the ability to perceive the distance of a simulated sound source. These include direct pointing methods (Parseihian et al., 2012, 2014), visual scale methods (Martin et al., 2021), verbal numerical reports (Kolarik et al., 2020) or cursor methods (Kopčo & Shinn-Cunningham, 2011).

In order to evaluate visual-to-auditory SSDs, the protocols used to assess distance perception abilities must be adaptable to blind people, provide measurements for near and far distances, and allow for a training phase to familiarize participants with the SSD encoding scheme through efficient audio-motor calibration. Given these constraints, visual and analogical scale methods or cursor methods can be excluded since their implementation requires functional vision. For their part, direct pointing or positioning methods have the disadvantage of being usable only for the evaluation of reachable distances. We therefore propose the "floor pointing method" as a new protocol for assessing absolute near- and far-field distance perception by measuring the distance indicated on the floor using a pointing tool.

The current study aimed to compare the early-stage abilities to perceive the distance of objects with an SSD as a function of the encoding scheme used for the distance dimension. Two encoding schemes were investigated: one using only intensity modulation and another which combined this with envelope amplitude modulation. To this end, distance localization performance using the two distance encoding schemes was assessed with a new type of pointing localization task, together with a distance discrimination task in a virtual environment in which the participants

were blindfolded. Near and far distances were used in the localization task to assess absolute distance perception. Only near distances were used in the discrimination task to assess relative distance perception.

2. Method

2.1 Participants

Eighteen participants (age: M = 24.17, SD = 4.76, 10 female, 16 right-handed) were enrolled in the study. No participant reported hearing impairments or any history of psychiatric illness or neurological disorder. The experimental protocol was approved by the local ethical committee Comité d'Ethique pour la Recherche de Université Bourgogne Franche-Comté (CERUBFC-2021-12-21-050) and followed the ethical guidelines of the Declaration of Helsinki. All the participants gave their written informed consent before the experiment, and they did not receive any monetary compensation.

2.2 Virtual environment

The experiment took place in a virtual environment created in UNITY3D software. The environment included the virtual targets to be localized and a virtual camera. The participant's head and a pointing tool, to which HTC VIVE Trackers 2.0 were attached, were tracked by four HTC VIVE base stations. The virtual environment could not be visually explored since the participants were blindfolded and were not wearing the virtual reality headset.

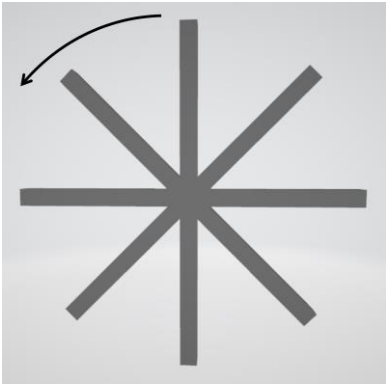
2.3 Virtual target

The virtual targets (see Figure 1) used in the experiment were 3D propellers composed of 4 intersecting bars as in Bordeau et al. (2023). In the familiarization sessions and the localization task, the length of the bars was 60 cm. In the discrimination task, 2 identical virtual targets with 5 cm-long bars were used. All the virtual targets rotated automatically at a speed of roughly 1.7 revolutions per second (10°/ 16.67 ms). Throughout the entire experiment, the participants could only perceive the virtual target through the SSD soundscapes.

2.4 Visual-to-auditory substitution

The visual-to-auditory substitution device used in the current study converts 3-dimensional spatial information (azimuth, elevation, and distance) contained in the processed video frames into soundscapes using the encoding scheme described in the following sections.

Figure 1. The virtual target to be localized was a geometric 3D propeller shape. The rotation direction of the target is indicated by the arrow.



2.4.1 Video processing

The virtual camera was linked to the head tracker attached to the participant's forehead. The position and rotation of the virtual camera were continuously updated on the basis of the position and rotation of the head tracker at a frame rate of 60 Hz. The field of view of the virtual camera was 90° × 74° (Horizontal × Vertical). The depth map of the image of the virtual environment captured by the virtual camera was encoded to form a 0-255 linear grayscale image from 0.01 m = 255 gray level (white) to 5.10 m = 0 gray level (black).

Video processing was performed in the same way as in Bordeau et al. (2023), where successive frame-differencing was achieved by computing the gray-level pixel-by-pixel absolute difference between the current frame and the previous one. Only pixels for which this difference was higher than 10 gray levels were kept in the image (the others were set to black). The black pixels are called "inactive" while the others are called "active" pixels (i.e., containing new visual information). Only the active pixels were converted into sounds and transmitted through the SSD soundscape. The processed frame was then scaled to a resolution of 160 x 120 pixels.

2.4.2 Visual-to-auditory conversion

The processed video stream was then converted into a stereophonic audio stream (or soundscape) which conveyed the extracted 3-dimensional information (azimuth, elevation, and distance) of each active pixel of the processed depth map in the form of acoustic information.

To this end, each of the 160 x 120 graphical pixels was associated with precomputed auditory pixels which took the form of a stereophonic sound containing acoustic cues relating to the position and the gray level of the graphical pixel it was associated with. Due to computational power limitations, auditory pixels were not computed in real-time. They were therefore pre-computed and stored in RAM at the beginning of the experiment. The encoding schemes were similar to those used in the *Monotonic condition* detailed in Bordeau et al. (2023).

2.4.2.1 Pitch modulation for the elevation encoding scheme

Each auditory pixel was computed based on a monophonic monotone of 35-ms duration with a 5-ms cosine fade-in and a 5-ms cosine fade-out. To convey the elevation, the frequency of the monotone was modulated from 250 Hz (bottom row) to 1492 Hz (top row) following the Mel scale.

2.4.2.2 Distance encoding scheme conditions

Two encoding schemes were tested for the distance dimension: one based on intensity modulation (INT encoding) and one combining intensity and envelope amplitude modulation (INT+ENV encoding). We pre-computed 26 layers of auditory pixels and each of them was associated with a distance range (from [0.01 m, 0.2 m] = layer #25 to [4.9 m, 5.1 m] = layer #0). The distance *d* associated with each pixel layer corresponded to the highest value of the interval (e.g., 0.20 m for layer #25). The distance encoding was applied to the monophonic monotones by multiplying them by a modulating function: $f_{INT}(d,t)$ for the INT encoding or $f_{INT+ENV}(d,t)$ for the INT+ENV encoding, where $t \in [0 ; 35 \text{ ms}]$ is the duration of an auditory pixel. Examples of monophonic tones for both distance encoding schemes are provided in Figure 2A.

2.4.2.2.1 Intensity modulation encoding

In the case of the Intensity encoding (INT), the distance encoding scheme is based on sound intensity modulation, which is a major acoustic cue in auditory distance perception. In anechoic environments, the attenuation of the sound intensity can be approximated to by the Inverse square law function $I(d) = \frac{I_0}{d^2}$, where *d* is the distance from the listener in meters (Blauert, 1983) and I_0 the sound intensity that would be measured at 1 meter from the sound source. To reproduce this acoustic cue in the current SSD, the amplitude of the auditory pixel was reduced with increasing distance. As previously mentioned, each gray pixel layer was associated with a distance *d* in meters. All the pixels of a given layer were thus modulated by the same multiplicative coefficient, which depended on the value of the associated distance *d*:

$$f_{INT}(s,d) = s \times R_{INT}(d),$$

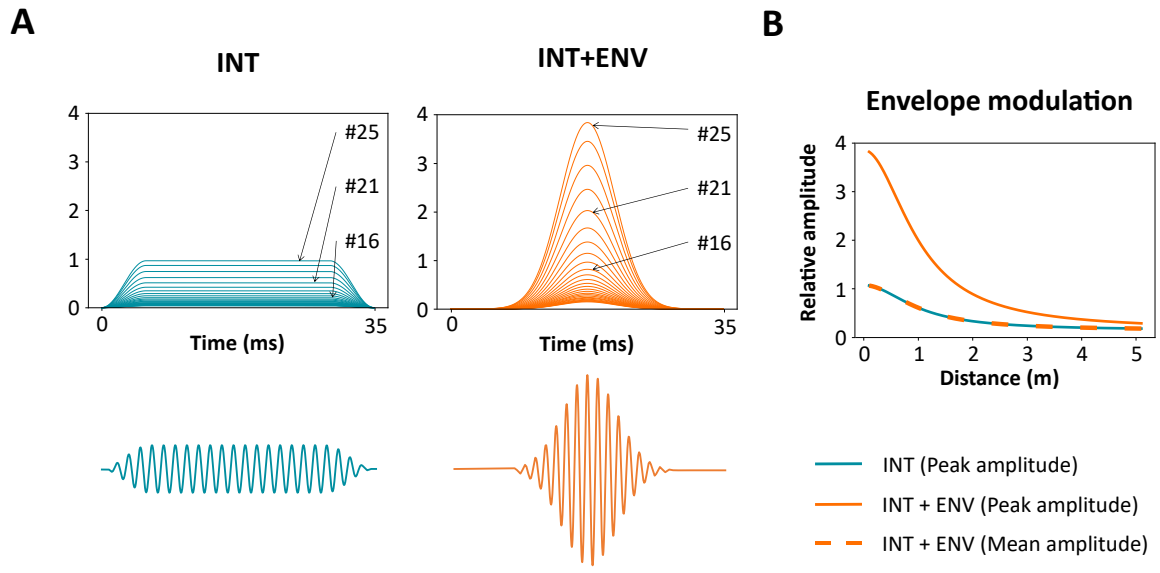
where *s* is the original 35-ms auditory pixel, and $R_{INT}(d) \in [0.04, 1]$ the amplitude modulation constant defined such as:

$$R_{INT}(d) = \frac{1}{c+d^2},$$

where $d \in [0.2; 5.1]$ is the distance in meters associated with the gray layer and $c = 1$ a constant. The amplitude modulation is depicted in Figure 2B. This function was chosen since the

intensity of the soundscape associated with a virtual target located at a distance *d* depended not only on the intensity modulation of individual auditory pixels, but also on the number of active pixels contained in the soundscape. Indeed, a closer target occupies a larger region of the camera's field of view, resulting in a higher number of active pixels, which are then contained in the soundscape. Due to the image processing method, the geometry of the target also influences the number of active pixels. Therefore, it would have been very hard to configure the increase in the intensity of the soundscape as a function of the proximity of the target based on the intensity modulation of the auditory pixel. The $R_{INT}(d)$ function was defined in order to approach the inverse square law while simultaneously limiting sound intensity for objects that came very close to the camera, thus preventing participants from experiencing upsettingly sounds of potentially upsetting intensity.

Figure 2. SSD distance encoding schemes. **(A)** The envelope amplitudes of the auditory pixels are depicted for the INT encoding (blue, left column) and the INT+ENV encoding (orange, right column) for the 26 gray layers (layers #25, #21 and #16 are highlighted as examples). An example of a monophonic tone (before the spatialization process) associated with the same location on the processed image and the same gray layer is provided for the two encodings. **(B)** Envelope amplitude modulation (peak amplitude and mean amplitude) of the auditory pixels as a function of the distance. The modulation of the peak amplitude is provided for the INT encoding (solid blue line) and the INT+ENV encoding (solid orange line), while the modulation of the mean amplitude is provided for the INT+ENV encoding (orange dashed line). Note that the modulation of the mean amplitude with the INT encoding is equivalent to the modulation of the peak amplitude.



2.4.2.2.2 Envelope and intensity modulation encoding

With the INT+ENV encoding, both intensity and envelope amplitude is modulated by a function of the following form:

$$f_{INT+ENV}(s,d) = s \times p(t) \times R_{INT+ENV}(d),$$

where *s* is the original 35-ms auditory pixel, and $d \in [0.2, 5.1]$ the distance in meters associated with the gray layer. All the auditory pixels of a gray layer were associated with the envelope amplitude modulation function *p(t)*, with $t \in [0, 35\ ms]$ as the time (i.e., duration of the auditory pixel), and an amplitude modulation constant $R_{INT+ENV}(d)$. The envelope was Gaussian-modulated with a shape of a normal probability density function:

$$p(t) = \left[\frac{1}{\sqrt{2\pi \times \sigma^2}} \times e^{\frac{1}{2} \times \left(\frac{t-\mu}{\sigma}\right)^2}\right],$$

where $t \in [0, 35\ ms]$ is the time, $\mu = 17.5$ ms (i.e., half of the duration of the auditory pixel), and $\sigma = 3.5$ ms. A Gaussian-modulated tone burst duration of more than 25 ms permitted pitch perception (i.e., just audible tonality) in the frequency range 250–1492 Hz (Mohlin, 2011). The function used to modulate the amplitude of the auditory pixels in the INT+ENV encoding was defined in such a way that the average amplitude of $f_{INT+ENV}(s,d)$ was equal to the average amplitude of $f_{INT}(s,d)$ for each gray layer.

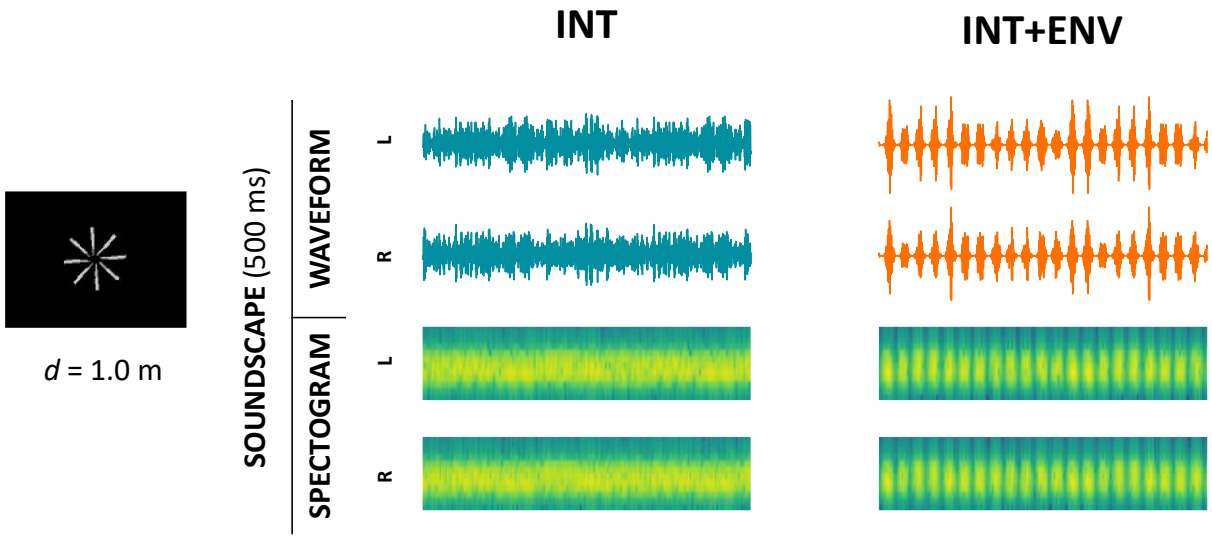
2.4.2.3 Sound spatialization for the azimuth and elevation encoding schemes

Each monophonic monotone was spatialized by convolving it with interpolated HRTFs from the CIPIC database (Algazi et al., 2001). These have a resolution of 5° in azimuth and 5.625° in elevation. The azimuth and elevation of the coordinates associated with each pixel corresponded to the projection of the position of the pixel within the image in the camera's field of view. A 4-point time-domain interpolation was computed to estimate the HRTFs to be applied. As in Sodnik et al. (2015), the Interaural Level Difference (ILD) and the convolution signals were separately interpolated using bilinear interpolations before being reassembled but using a 2D interpolation instead of a 1D interpolation.

2.4.2.4 Audio mixing for soundscape generation

All auditory pixels of each new video frame were summed to form an audio frame. In order to generate a smooth audio stream, each new computed audio frame was added to the end of the previous one with an overlap lasting 5ms corresponding to the 5-ms cosine fade-in and fade-out of the auditory pixels. Examples of soundscapes for the two distance encodings are given in Figure 3.

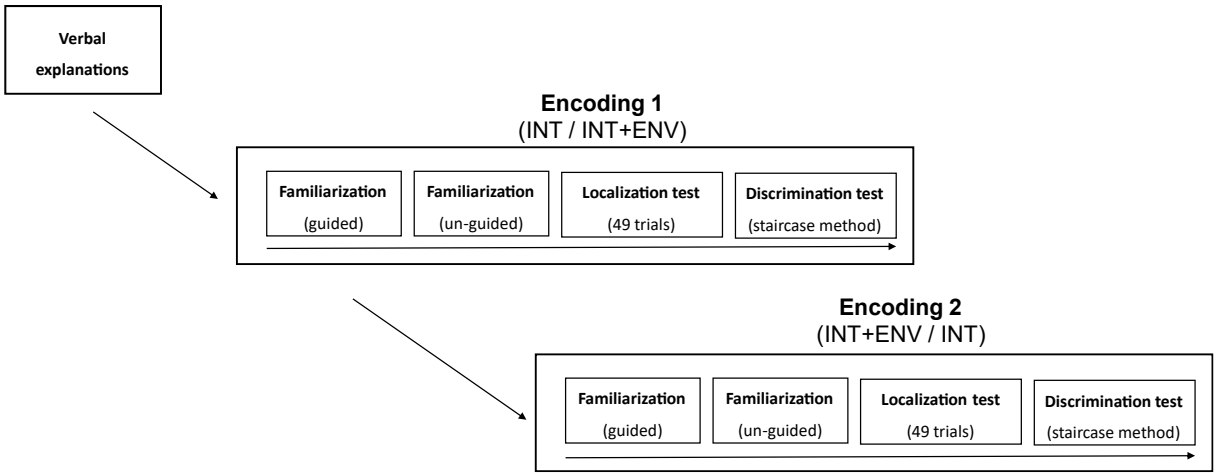
Figure 3. SSD soundscapes. Example of a video frame captured by the virtual camera in which a target is located at d = 1.0 m from the virtual camera. The waveform and spectrogram of the associated soundscape are provided for the INT encoding (left column) and the INT+ENV encoding (right column), and for the left (L) and right (R) ear channels separately.



2.5 Experimental Procedure

The experiment consisted of a 1-hour session during which participants sat comfortably in a chair located in a room surrounded by the virtual reality tracking system. The participants were equipped with a SONY MDR-7506 headphone which was used to deliver the soundscapes. The timeline of the experimental session is depicted in Figure 4.

Figure 4. Timeline of the experiment. After the experimenter had given verbal explanations about the principle of the SSD, the participant tested the two encodings one after the other. For each encoding, the participant experienced both guided and un-guided familiarization and then performed the localization task and the discrimination task.



At the beginning of the experiment, the experimenter told the participants that they would have to localize a virtual target located in front of them at different distances on the floor by

pointing to it. Participants were instructed to remain seated on the chair during the experiment. They were blindfolded except during the guided familiarization phase and breaks.

The experimenter started by giving a verbal explanation of the main principles of the visual-to-auditory SSD used in both tested encodings. After confirming that they had understood the explanations and agreed to continue, each participant started to test one of the two visual-to-auditory distance encodings (INT and INT+ENV encodings). The order of the two tested encodings was counterbalanced between participants. For each encoding and before the localization tasks began, the participants were familiarized with the visual-to-auditory encoding by means of a guided familiarization followed by an unguided familiarization phase. After the familiarization session, participants performed the distance localization task followed by the distance discrimination task. These steps are detailed in the following sections.

2.5.1 Verbal explanations of the main principles of the SSD

At the beginning of the experiment, the experimenter gave verbal explanations of the main principles of the SSD. The experimenter explained to the participants that they would not be able to see the virtual target but would instead only be able to hear it and that the sound would depend on the position of the target relative to the position and the direction of their head. The experimenter explained the main principles shared by the encoding schemes: spatialization for the azimuth, pitch modulation for the elevation and intensity modulation for the distance. Although participants were told that two different encoding schemes would be tested, no information regarding their specific differences was given.

2.5.2 Pointing tool

A pointing tool was used in the familiarization sessions and localization task. The pointing tool was a tracked pistol. Participants were either instructed to use it to position the target on the floor in the familiarization sessions, or to indicate the perceived target position in the localization task. They were instructed to hold the pointing tool with their dominant hand with arm outstretched.

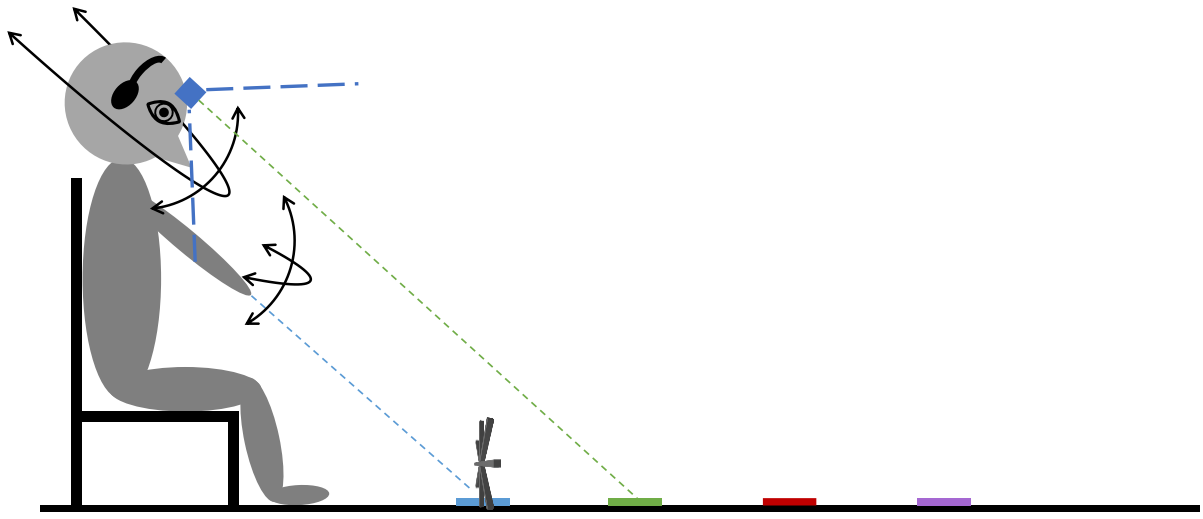
2.5.3 Guided and unguided familiarization

The familiarization session was conducted to help participants learn how the sound was modulated depending on the position of the target relative to their head orientation (i.e., relative to the virtual camera's field of view). The familiarization session was first performed sighted and guided before then being undertaken blindfolded and unguided. In both sessions, the participant used the pointing tool to control where the target was placed on the floor. The position of the

target was continuously updated as the intersection of a virtual ray originating at the tip of the pointing tool with the virtual floor. The x-coordinate (lateral position), y-coordinate (vertical position) and z-coordinate (distance position) of the target were continuously monitored. The position and rotation of the virtual camera were continuously updated based on the position and the rotation of the head tracker.

The first part of the familiarization phase consisted of sighted guided familiarization, comprising three successive rounds during which the experimenter told the participants where to point on the floor with the gun in order to position the target and towards which of 4 colored crosses on the floor they should direct their heads (Figure 5). This guided familiarization session was conducted so that participants could actively familiarize themselves with the encoding scheme and understand that both the orientation of their heads and the location of the target modified the soundscape. In the first round, participants moved the target while keeping their heads facing in the direction of a cross. In the second, they moved their heads while maintaining the target positioned at a cross. In the last round, they simultaneously moved both target and head toward the same cross. All participants received the same instructions in the same order.

Figure 5. Guided familiarization method. The non-blindfolded participants had to sequentially position the target (gray propeller) and direct their heads toward different crosses on the floor. The experimenter told the participants which of the 4 crosses (blue, green, red and purple lines) they had to direct their heads towards and where they had to position the target.



After the sighted guided familiarization session, participants practiced unguided familiarization, which consisted of a 60-sec session during which they were free to move their heads and position the target wherever they wanted in order to actively explore the way this modified the soundscape.

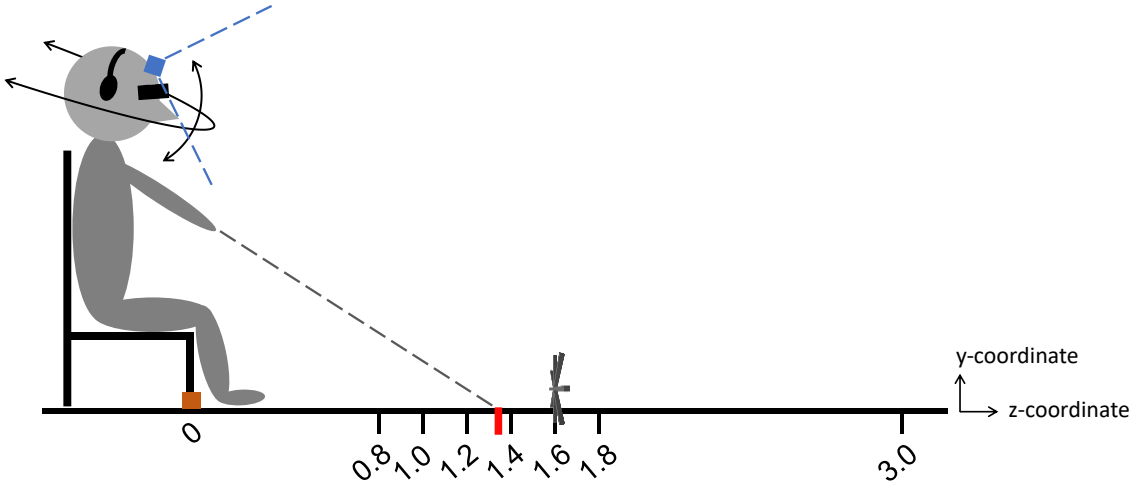
2.5.4 Reference point in the virtual environment

During the absolute distance localization task, the positions of the target were defined relative to a reference point that was located on the floor under the chair on which the participant was sitting (Figure 6). The location of this reference point was computed once at the beginning of the experiment using a Vive tracker placed on the floor.

2.5.5 Localization task

Absolute distance perception abilities were assessed with a localization task using the "floor pointing method". The localization task consisted of 49 trials during which blindfolded participants had to localize the virtual target based on the soundscapes provided by the visual-to-auditory SSD. During each localization task, the target was positioned on the floor at one of seven possible distances (0.80, 1.00, 1.20, 1.40, 1.60, 1.80 and 3.0 m) from a reference point located on the floor under the participant. Figure 6 illustrates the experimental set-up. The order of the tested positions was randomized, and each position was tested 7 times during each of the two localization tasks (i.e., one localization task per encoding).

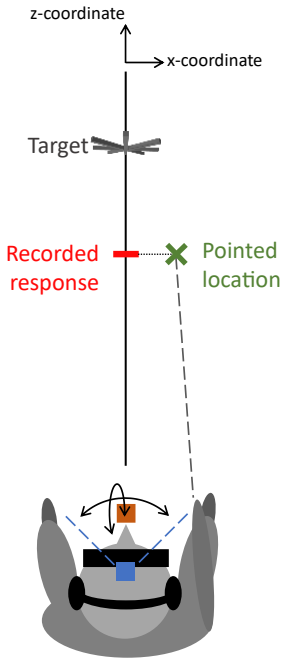
Figure 6. Localization task. The blindfolded participant had to localize a target on the floor (gray propeller located at 1.6 m) by pointing to it (red vertical line). The target could be positioned at one of 7 locations ranging from 0.80 to 3.00 m from the reference point located under the participant (brown square). The virtual camera position and orientation were continuously updated based on the participant's head tracker (blue square).



Each trial started with a 500-ms 440-Hz beep which indicated the beginning of the trial. After a 500-ms silent period, the virtual target was displayed at one of the 7 tested positions. The orientation of the target was controlled so that it faced the virtual camera. Participants were instructed to hold the pointing tool in an outstretched arm and point to the perceived location of the target. They were allowed and encouraged to move their heads in order to localize the target. No time limit was imposed for responding but participants were asked to respond as fast and

accurately as possible, while prioritizing accuracy. The virtual target was displayed until participants pressed the trigger of the pointing tool, causing the response position to be saved. The perceived distance was computed as the distance between the reference point and the z-coordinate of the intersection point (Figure 7). After a 1000-ms inter-trial break, the 500-ms 440-Hz beep sounded again to indicate the start of the next trial. No feedback regarding response accuracy was provided.

Figure 7. Response collection method. In the localization task, the participant had to localize a target (gray propeller) by pointing to it with the pointing tool. The recorded distance response (red line) was computed as the z-coordinate of the pointed location (green cross) on the virtual floor.

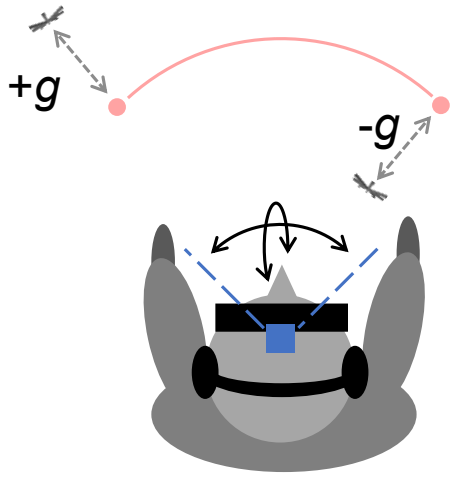


2.5.6 Discrimination task

In order to assess relative distance perception, a discrimination task similar to the one used by Richardson et al. (2019) was conducted. The discrimination task (Figure 8) consisted of successive trials during which the blindfolded participants had to verbally report to the experimenter which side the closest of two simultaneously presented targets was on (two-alternative forced-choice method). Each trial started with a 500-ms 440-Hz beep to indicate the beginning of the trial. After a 500-ms silent period, the two virtual targets were simultaneously displayed at different distances from the participant but at the same height as the position of the head tracker at the time of appearance. During each trial, one of the targets was located on the left side of the participant (-40° azimuth) and the other on the right side (40° azimuth). Participants were instructed to verbally report to the experimenter which target was the closest ("left" or "right"). The location of the nearest target was randomized between trials. Participants were allowed and encouraged to move their heads to localize the targets and no time limit was imposed for responding. The virtual targets were displayed until the experimenter recorded the participant's

response in the Unity interface. After a 1000-ms inter-trial break, the 500-ms beep sounded to indicate the start of the next trial. The two virtual targets were displayed at a distance of 80 ± *g* cm (2*g* being the gap between the two targets). The distance *g* changed in each trial following a 3-down/2-up staircase method. The initial tested distance *g* was 50 cm. Each wrong response caused *g* to increase by +Δ cm, and each correct response caused it to decrease by -Δ cm, with Δ being set to 5 cm at the beginning of the experiment. The Δ value was divided by two on each wrong answer. The experiment stopped after three wrong responses had been given, and the discrimination score was computed as the mean value of *g* for the last two trials. No feedback was provided regarding the correct response and the distance *g* was automatically updated in accordance with the staircase protocol.

Figure 8. Discrimination task. The blindfolded participant had to indicate which of two targets (gray propellers) was the closest. Targets were positioned at 80 cm +/- g from the participant. A staircase method was used to compute the discrimination score, with g = 50 cm at the beginning of the experiment. g increased by Δ on each wrong response and decreased by -Δ on each correct response. The Δ value was 5 cm at the beginning of the experiment and was divided by two on each wrong answer (i.e., taking the following values {-5, +2.5, -2.5, +1.25, -1.25} cm). The virtual camera position and orientation were continuously updated based on the data from the participant's head tracker (blue square).



2.6 Data analysis

Statistical analyses were performed using R (version 3.6.1) (Team, 2020). Localization performance in the pointing localization task was assessed both with regression-based metrics and error-based metrics, while performance in the discrimination task was assessed on the basis of the discrimination score. The *lmerTest* R-package (Kuznetsova et al., 2017) was used to fit the data with Linear mixed models (LMMs), and version 1.7.4 of the *emmeans* R-package (Length, 2022) and Tukey HSD correction were used to conduct post-hoc analyses.

2.6.1 Data analysis of the localization task with regression-based metrics

Data from the localization task of one subject (S05) were discarded since the subject did not understand the instructions given by the experimenter. Outlier removal was performed with a two-step process. First, distance response locations that were higher than 10 m were considered as outliers. The second step then consisted in removing the distance response locations that were outside the range [M ± 3 × SE] = [-1.66 m, 4.41 m]. The two steps resulted in the deletion of 35 trials in total, corresponding to the deletion of 2.1% of the raw data.

Power functions of the form $d' = k \times d^a$ are known to provide good approximations of the psychophysical function that relates the estimated perceived distance to physical sound source distance (Zahorik, 2002, 2005), where d' is the estimate of the perceived distance, and d the true target distance. k and a are the parameters of the fitted power functions. These power functions become linear when relating the logarithmically transformed perceived distances to the logarithmically transformed true distances and they take the form: $d'_{log} = A \times d_{log} + B$, where d'_{log} is the predicted log-transformed distance response, d_{log} the log-transformed true target distance, A = a the slope of the function, and B = log(k) a constant. The data from the localization task were therefore logarithmically transformed before being fitted by means of an LMM. The log-transformed response positions by participants and by trials were included in the LMM. The model included the participants as random factor and Encoding (INT or INT+ENV) and Target-distance (log-transformed) as fixed factors.

The predictions from the LMM were used to approximate to the exponential values a (slope of the model) and constant k (the exponentially transformed intercept of the model). An exponential value a equal to 1.0 and a constant k equal to 1.0 would correspond to optimum localization performance. An exponential value a below 1.0 indicates a distance compression bias (i.e., an overestimation of nearby distances and an underestimation of far distances) while, on the contrary, an a value larger than 1.0 indicates a distance extension bias. The estimated constant k provides an insight into the veridical distance, which is the distance at which a switchover is observed between the overestimation and underestimation pattern (i.e., where the distance is best reported). The effects were estimated using an ANOVA. Post-hoc analyses with Tukey HSD correction were conducted whenever a significant effect was found and two-tailed paired t-tests were used to compare the exponent values a between the INT and INT+ENV distance encoding schemes. One-sample t-tests were used to demonstrate the deviation of the exponent value a from the optimum value 1.0.

2.6.2 Data analysis of the localization task with error-based metrics

The unsigned error was computed as the absolute difference between the recorded distance response position and the target distance position for each trial. Unsigned errors were fitted by means of an LMM which included the participants as random factor and Encoding (INT or INT+ENV) and Target-distance as fixed factors. The effects of the fixed factors were estimated using an ANOVA. Post-hoc analyses with Tukey HSD correction were conducted whenever a significant effect was found and two-tailed *t*-tests were used to compare the estimated marginal means of the unsigned error between the conditions.

2.6.3 Data analysis of the discrimination task

Data from two subjects (S01 and S13) were discarded due to a technical issue during the experiment. No outlier removal was conducted. A two-tailed *t*-test was used to compare the discrimination score between the encodings (INT or INT+ENV).

3. Results

3.1 Head tracker checks during the pointing task

As previously mentioned, the target locations in the pointing task were set relative to a reference point located on the floor (Figure 6). However, as explained in the Method section, the sonification of the target through the virtual SSD was performed relative to the camera position and not relative to this reference point. For descriptive purpose, Table 1 reports the average distances measured between the participant's head and the target, as a function of the distance between the reference point and the target.

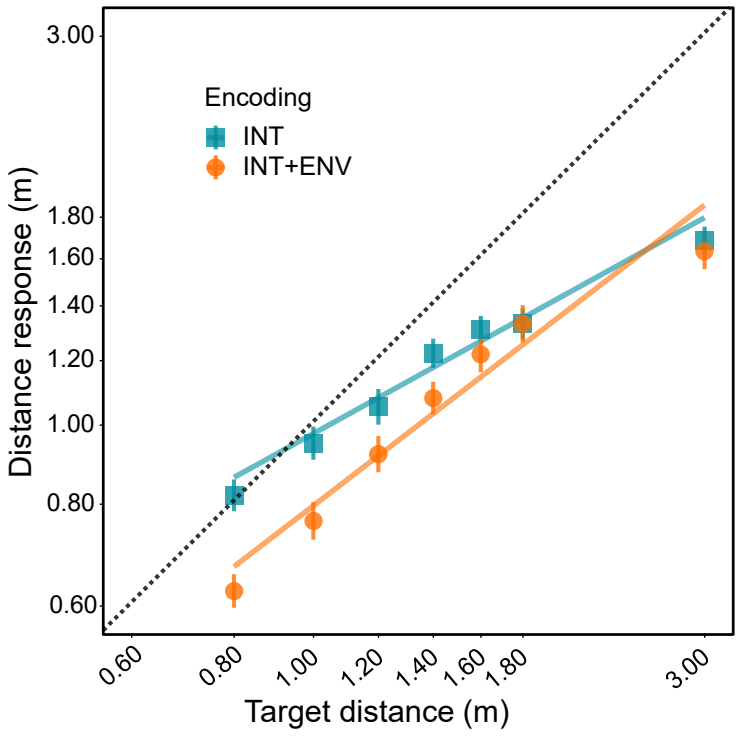
Table 1. Distance between the target and the participant's head at the time of the response recording during the localization tasks for each tested distance. Note that participants could freely move their heads during each trial, with the result that the distance between the target and the participant varied throughout each trial.

Target distance (m) distance between the reference point and the target	**Distance (m) between the head tracker and the target** (mean ± standard deviation)
0.80	1.31 ± 0.11
1.00	1.46 ± 0.11
1.20	1.62 ± 0.10
1.40	1.79 ± 0.10
1.60	1.96 ± 0.10
1.80	2.14 ± 0.11
3.00	3.26 ± 0.09

3.2 Distance localization performance in the pointing task

Perceived distances are depicted using logarithmic scales in Figure 9. The slope of the linear fit of the log-transformed data is equivalent to parameter a of the power function, and the intercept of the linear fit of the log-transformed data is equivalent to log(k) (see details in the section 2.6.1 Data analysis of the localization task with regression-based metrics). The ANOVA revealed a significant interaction effect between Encoding and Target-distance [$F(1,15.13) = 9.11$, $p = 0.0086$, $\eta_p^2 = 0.38$], suggesting that the response patterns were different for the INT and INT+ENV distance encodings. Post-hoc comparisons of the slopes revealed a compressive bias with both encodings. With the INT encoding, the exponent value a was 0.555 and was significantly lower than 1.0 [$t(16) = 7.472$, $p < 0.0001$]. The value of a was 0.773 with the INT+ENV encoding and was also significantly lower than 1.0 [$t(16) = 3.177$, $p = 0.0059$]. With this encoding method, this exponent value was found to be significantly higher than the exponent value with the INT encoding [$t(16) = 3.017$, $p = 0.0082$], suggesting that the compressive bias is lower with the INT+ENV encoding scheme.

Figure 9. Mean distance response position (in m) as a function of target distance for the INT (blue squares) and INT+ENV (orange circles) encodings in log coordinates. Solid lines represent the slope of the linear model on log coordinates (i.e., the a exponent value in the power functions). Error bars show standard errors. Black dashed lines represent the optimal exponent value a = 1.0 and constant value k = 1.0.



The value of the constant k was approximated to by computing the exponential value of the intercept of the model with the log coordinates. With the INT encoding, the intercept was -

0.02 (95% CI = [-0.19, 0.14]), which corresponds to a value of k = 0.98 (95% CI = [0.83, 1.15]). This intercept was measured at -0.23 with the INT+ENV encoding (95% CI = [-0.43, -0.02]), which corresponds to a value of k = 0.79 (95% CI = [0.65, 0.98]). In other words, the veridical distance (i.e., the distance to the target at which the overestimation response pattern switches to an underestimation pattern) was 0.98 m with the INT encoding and 0.79 m with the INT+ENV encoding.

Localization performance was also assessed with the distance unsigned error. Figure 10 shows the unsigned error as a function of target distance. The ANOVA revealed a significant effect of Target-distance on the unsigned error [$F(6, 22.15) = 9.868$, $p < 0.0001$, $\eta_p^2 = 0.73$], while there was no significant effect of Encoding and no significant interaction between Encoding and Target-Distance. Table 2 summarizes the distance unsigned errors (with 95% confidence interval) and the results of the two-tailed *t*-test comparisons between the Target-distance conditions. Comparisons showed a general trend for the unsigned error to increase with increasing distance and this finding was consistent among the farthest distances, although the gap of 0.2 m between each target distance did not systematically result in a significant increase of the unsigned error at smaller distances (see Table 2). Overall, target distance localization accuracy tended to decrease with increasing distance independently of the employed encoding scheme.

Figure 10. Estimated marginal mean of the distance unsigned error (in m) as a function of target distance. Error bars show the 95% confidence interval.

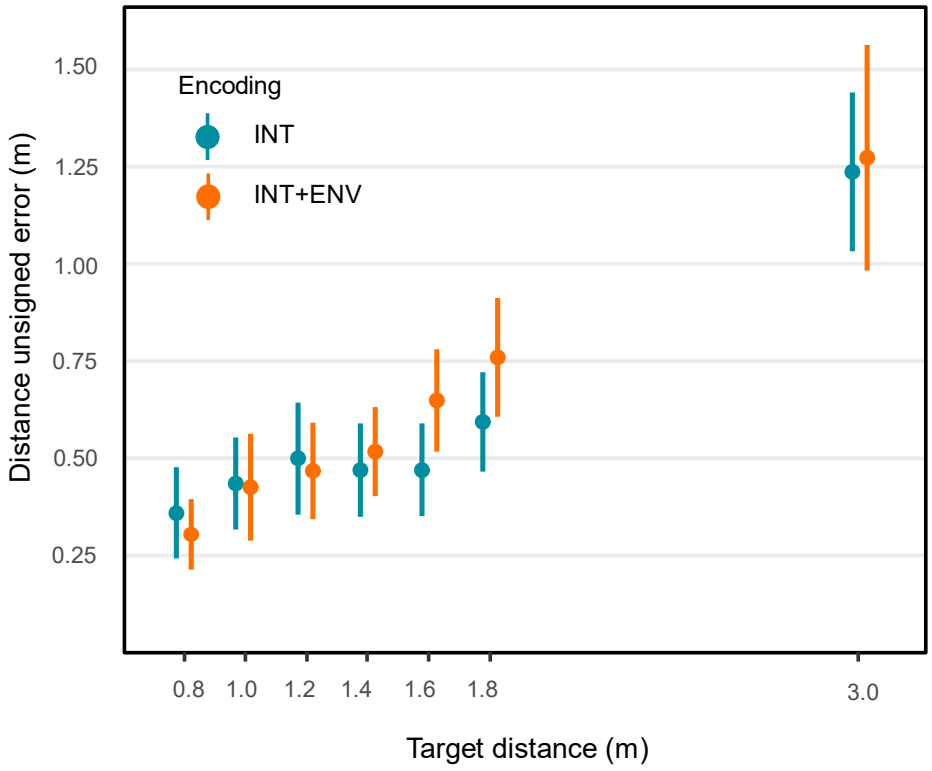


Table 2. Post-hoc comparisons of the distance unsigned error between the target distance positions. The estimated marginal means for each target distance are depicted along the diagonal, with 95% confidence interval in brackets. When a significant difference between two target distances was found, the p-value is provided, while ns. is reported if no significant difference was found.

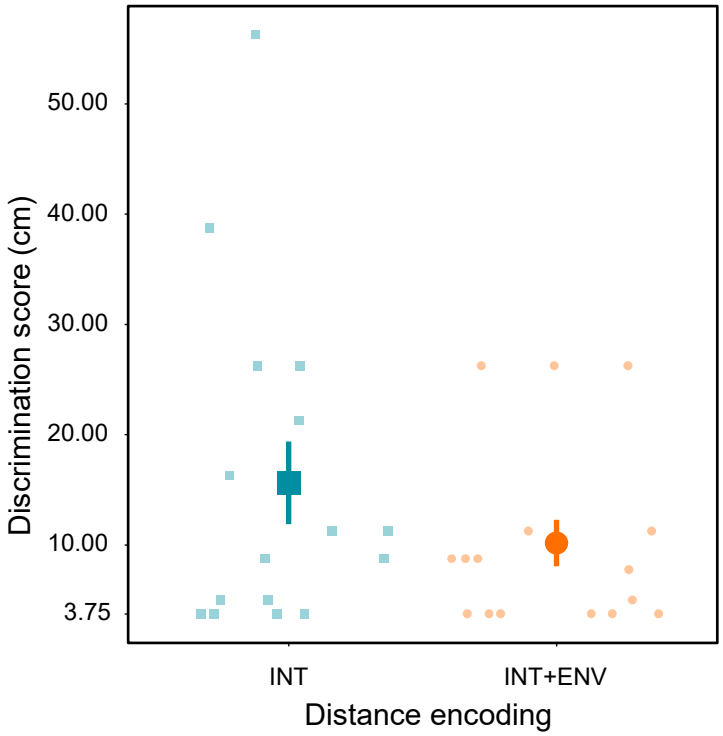
Target distance (m)	**0.80**	**1.00**	**1.20**	**1.40**	**1.60**	**1.80**	**3.00**
0.80	0.33 m [0.24, 0.42]	ns.	$p = 0.0371$	ns.	$p = 0.0283$	$p = 0.0004$	$p < 0.0001$
1.00	.	0.43 m [0.32, 0.54]	ns.	ns.	ns.	$p = 0.0092$	$p = 0.0001$
1.20	.	.	0.48 m [0.36, 0.61]	ns.	ns.	$p = 0.0221$	$p = 0.0002$
1.40	.	.	.	0.49 m [0.39, 0.59]	ns.	$p = 0.0066$	$p = 0.0001$
1.60	.	.	.	.	0.56 m [0.44, 0.68]	ns.	$p = 0.0001$
1.80	.	.	.	.	.	0.68 m [0.56, 0.80]	$p = 0.0005$
3.00	.	.	.	.	.	.	1.26 m [1.02, 1.50]

3.3 Distance discrimination performance

Discrimination scores for both encodings are depicted in Figure 11. The discrimination score with the INT encoding was 15.63 ± 14.98 cm (M ± SD). With the INT+ENV encoding, the discrimination score was 10.17 ± 8.42 cm (M ± SD), which was significantly lower than the discrimination score with the INT encoding [$t(15) = 2.178$, $p = 0.046$, $d = 0.545$]. More than half of the participants (9 out of 16) obtained a lower discrimination score with the INT+ENV encoding, 4 had a lower discrimination score with the INT encoding, and 3 participants performed equally with both encodings.

However, the results also suggest a ceiling effect. Due to the staircase protocol, the lowest discrimination score that could be obtained was 3.75 cm. Over the 16 participants, this score was observed 4 times with the INT encoding and 6 times with the INT+ENV encoding. Although the ceiling effect suggests that the discrimination performances could have been improved for some participants by conducting a longer experiment (especially with the INT+ENV encoding), the results of the discrimination task suggest that envelope amplitude modulation had a facilitating effect for relative distance perception with the SSD within the nearby space.

Figure 11. Mean discrimination scores (cm) with the INT encoding (large blue square) and the INT+ENV encoding (large orange circle). Error bars show the standard error. Individual discrimination scores of the 16 participants with both encoding methods are depicted in light blue squares and light orange circles, suggesting a ceiling effect.



4. Discussion

The aim of the current study was to compare two encoding methods implemented in a visual-to-auditory sensory substitution device (SSD) in order to convey spatial information about the distance: sound intensity modulation only or sound intensity coupled with envelope amplitude modulation. Distance perception abilities were evaluated using absolute and relative distance perception tasks performed by blindfolded participants. Absolute distance perception was assessed using a new pointing method, while relative distance perception was assessed with a distance discrimination task. Before performing the tasks, participants were familiarized with the encoding schemes through an active audio-motor familiarization session.

4.1 Sound envelope amplitude modulation can reduce the compressive bias

A floor pointing localization task was used to assess the participants' ability to perceive the absolute distance of the virtual object with the SSD. Log-transformed data were linearly fitted to estimate the exponent value a and the constant k of the power function $d' = k \times d^{a}$, where d is the physical distance and d' is the estimated distance predicted by the model. A compressive bias is indicated by exponent values a lower than 1.0, which reflect a tendency to overestimate nearby distances and underestimate far distances. The estimated constant k provides information about

the magnitude of the overestimation of nearby distances by giving an approximation to the veridical distance, which is the distance where distances are best reported.

In our study, a compressive bias was observed both with the distance encoding scheme which modulated sound intensity (a = 0.55) and the scheme which combined sound intensity and envelope amplitude modulation (a = 0.773). However, this bias was lower with the latter encoding scheme. When encoding was performed by modulating both sound intensity and envelope amplitude, the veridical distance was 0.79 m, whereas this value was equal to 0.98 m when encoding was performed using only sound intensity. Overall, the judgment of distance with the encoding scheme which combined sound intensity and envelope amplitude modulation was less compressed than with sound intensity modulation only, suggesting that distances, and in particular nearby distances, are estimated better with this encoding scheme.

The compressive perception of distance has been reported in many studies on distance perception of simulated sound sources (Bronkhorst & Houtgast, 1999; Kolarik et al., 2020; Kopčo & Shinn-Cunningham, 2011; Martin et al., 2021; Zahorik, 2002) and real sound sources (Parseihian et al., 2014, see Kolarik et al., 2016; Zahorik, 2005 for review). These related works are discussed in more detail in the following sections.

4.1.1 Distance compressive bias with real and simulated sound sources using individualized HRTFs

The exponent value a and the constant k measured in our study are consistent with the averaged values of 33 studies on simulated or real sound source distance localization reported by Zahorik (2002). Although there are differences between the studies, the compressive bias was systematically found, with an average exponent value a of 0.59 ± 0.24 (M ± SD) and constant k of 1.66 ± 0.92 (M ± SD). In our study, the estimated a and k parameters with both encodings were within 1 standard deviation of the average estimates reported in Zahorik (2002), despite the diversity of the types of sound (real or simulated with individualized HRTFs), tested distance ranges, employed reporting methods, and visual contexts (eyes open or closed), all of which are known to influence the measurements (Zahorik, 2005).

In their studies, Bronkhorst & Houtgast (1999) and Kopčo & Shinn-Cunningham (2011) used simulated sound sources and distance ranges comparable to the distance range used in our study. Bronkhorst & Houtgast (1999) tested distances from 0 to 3.5 m and measured an exponent value a of 0.44, while Kopčo & Shinn-Cunningham (2011) tested a slightly shorter distance range of between 0.15 to 1.7 m (but still comparable to the one used in our study) and measured an

exponent value *a* of 0.65. Their estimates are comparable to the exponent values *a* of 0.55 and 0.773 measured in the current study, despite the use of individualized HRTFs.

Kopčo & Shinn-Cunningham (2011) also used simulated sound sources with individualized HRTFs and focused on how the frequency spectrum of the simulated sound source modulates distance perception abilities. They measured the effect of the frequency composition and bandwidth of the frequency spectrum on the exponent value *a* and also measured the highest exponent values (about 0.65) when a broadband sound (300–5700 Hz) or a wideband low-pass filtered sound was used (300–3000 Hz). Conversely, a stronger compressive bias with exponent values of about 0.55 was measured when a narrowband low-pass filtered sound was used (300–500 Hz). In our study, the soundscapes could contain energy in the frequency range between 250 and 1492 Hz. Therefore, the frequency spectrum of the soundscapes could result in different distance localization abilities depending on the elevation of the target in the camera's field of view.

4.1.2 Distance compressive bias with simulated sound sources using non-individualized HRTFs

The above-mentioned studies used individualized HRTFs recorded at different distances to simulate the sound sources, while non-individualized HRTFs recorded at a single distance were used in the current study (from the CIPIC database, Algazi et al., 2001). Since the database contains HRTFs recorded at 1 m, the SSD distance encoding scheme was applied to tones that were subsequently spatialized with HRTFs corresponding to a unique distance.

Martin et al. (2021) compared spatialization rendering methods used to simulate sounds at various distances and showed that distance localization abilities are influenced by the rendering method. The "measurements-based" rendering method used a Binaural Room Impulse Responses (BRIRs) dataset recorded at 9 distinct distances (from 1 to 7 m), whereas the "intensity-based" rendering method used a single BRIRs dataset (1 m) to which intensity modulation was then applied to reproduce the way in which intensity varies with changes in distance. The "intensity-based" method can therefore be compared to the distance encoding schemes that we used in the SSD. In a similar way to in the current study, Martin et al. (2021) measured a compression bias with an exponent value a lower than 1.0 with this method, whereas distance localization performance was higher when the sound source was simulated with the "measurement-based" rendering method. We might therefore expect to observe improved distance localization abilities with the current SSD by using HRTFs recorded at different distances as in the "measurement-based" method in Martin et al. (2021).

However, the distance perception of simulated sound sources can also be drastically impaired when non-individualized HRTFs recorded at different distances are used. For instance, Parseihian et al. (2014) used a blindfolded pointing task within the reachable space (from 0.33 to 0.85 m) and observed that participants were not able to perceive distance changes with the employed non-individualized HRTFs since they consistently judged the distance to correspond to the same location (slope A = -0.03, in linear coordinates). In the same study but with real sound sources, the distance was perceived even though a strong compression bias was observed with an average slope A = 0.25. However, in contrast to the current study, the sound source was simulated only at very close distances. The difficulty of perceiving the distance of simulated sound sources with non-individualized HRTFs indicates the importance of integrating additional acoustic cues to convey distance in the context of SSD in order to improve the perception of nearby obstacles and limit the risk of collision.

4.1.3 Distance compressive bias with sensory substitution devices

The distance compression bias has also been reported in the context of visual-to-auditory SSD, for example in Commere & Rouat (2023) and Parseihian et al. (2012) who used, respectively, repositioning or direct pointing methods within the reachable space. Both studies also found a compression bias even though they reported the estimated distances with a linear relation with respect to the true distances instead of a power fitting. Using the repositioning method for distances ranging from 0 to 0.90 m, Commere and Rouat (2023) found a compression bias for all of the five studied encodings: the modulation of amplitude, reverberation, pitch, beep repetition rate, and signal-to-noise ratio. Parseihien et al. (2012) used a direct pointing task to compare three encoding schemes intended to convey distances ranging from 0.73 to 1.07 m. The authors obtained a compression with both signal-to-noise ratio encoding and beep repetition rate encoding with slopes A of 0.57 and 0.96. However, using a third encoding scheme based on reverberation, distance was not perceived at all. Whereas veridical distances (i.e. the distance at which the overestimation pattern switches to the underestimation pattern) are often found around 1.5 m Zahorik (2005), it was measured at 0.50 m on average in the above study.

Due to geometrical considerations, increasing the distance of the target on the floor while keeping the head still resulted in the target being located higher in the virtual camera's field of view in our study. Since the employed SSD uses pitch modulation (from 250 to 1492 Hz) to convey elevation, a target at a higher location produces a soundscape with higher-frequency components. In this case, moving the target on the floor away from the participant results in both intensity and pitch changes in the soundscape. It is thus possible that some of the participants used both distance encoding and elevation encoding to localize the target.

In addition to a compressive bias, we observed a decrease in accuracy as the target distance increased (unsigned error from 0.33 m for the closest distance to 1.26 m for the farthest distance). Auvray et al. (2007) used a direct pointing task to investigate the ability to localize a real object with the vOICe SSD (Meijer, 1992) and found a mean pointing error of 7.8 ± 5.1 cm, which increased with the distance from the participant. However, it is not easy to compare our results with theirs since, in their study, the object was located on a table within the reachable space and the lateral error was comprised in the measured pointing error.

Overall, our results in the localization task are in line with previous results showing a distance compression bias in both auditory localization experiments and SSD experiments. However, the results suggest that envelope amplitude modulation has a facilitating effect on absolute distance localization abilities, with both the compression bias and the overestimation of near distances decreasing. It is important to reduce the overestimation of near distances in the context of SSD use during pedestrian trips since this can help reduce the likelihood of colliding with obstacles.

The tendency for the localization error to increase with the target distance was similarly observed with both encodings. In the context of SSD, the accurate distance estimation of close objects is a priority. The facilitation effect of the envelope amplitude modulation that we observed is therefore promising since this encoding can be easily implemented on SSD for the sonification of depth maps and can result in better distance judgments. Since the loudness of a soundscape does not only depend on its intensity, it is still necessary to assess the perceptive modulation of the loudness with each distance encoding scheme. In the context of SSD, the intensity of a soundscape corresponding to a target depends on many parameters, such as the proximity of the target, its location within the camera's field of view and its geometry. This gives rise to methodological limitations when assessing SSD performances in real use since there is considerable variability in the displayed soundscapes when the participant's head is not restrained. At the same time, this also gives users a great opportunity to develop complex exploration strategies. Despite these methodological considerations, it appears essential to conduct experiments in contexts that are closer to real-life situations, while also considering the resulting inter-individual and inter-trial variability.

4.1.4 The floor pointing method: a new way to assess absolute distance perception

The current study proposes a new method called "the floor pointing method" to assess absolute distance perception of sound sources. This method possesses a unique combination of three characteristics that make it well-suited for absolute distance measurements in the context of SSD evaluation. First, the floor pointing method makes it possible to assess distance perception

abilities in a range of distances that are relevant in SSD use. The second characteristic is the possibility of administering training to allow users to familiarize themselves with the sensorimotor contingencies by associating motor actions (i.e., head movements and arm directions with the pointing tool) with changes in the sensory stimulation (i.e., the soundscape) (Auvray, 2004). The third characteristic, which is of the utmost importance, is the possibility of adapting this method for use by blind people, who are, ultimately, the targeted population.

Previous studies suggest that blind people are better able to judge the relative distance of sound sources but less able to perceive their absolute distance (Kolarik et al., 2016). Although supranormal abilities have been measured for relative distance perception in the blind (Kolarik et al., 2013; Voss et al., 2004), these have not been systematically observed. For instance, the distance compression bias has been shown to be more pronounced in blind than in sighted participants (Kolarik et al., 2017), and absolute distance perception can be slightly more impaired in the blind in a pointing task within the reachable space (Macé et al., 2012). Therefore, the ability to perceive distance with the SSD encoding used in the current study remains to be assessed in blind people. For this reason, the experimental procedure was designed in a way that is reproductible in blind people. The only exception was the guided familiarization phase, which can easily be adapted.

The errors we measured in the pointing task have two main origins: perceptual errors in the localization of the sound source itself, and proprioceptive bias when using the pointing tool to point towards a specific position while being blindfolded. Obvious geometrical considerations show that, with our method, pointing at far distances intrinsically results in less precision since a small angular error induces a higher difference in the distance measured to the indicated point on the floor when pointing at more distant rather than nearby targets. This effect could be particularly salient since the blindfolded participants in this study could only rely on proprioceptive cues to point at the target given that they received no visual feedback. This effect is probably the cause of the increase in the standard deviation of the distance error from 0.39 m for the nearest to 0.95 m for the farthest position (Table 2) and may also be reflected in the presence of distance response positions recorded at more than 10 m (18 discarded outliers). However, a study that focused more specifically on this topic suggests that proprioceptive cues are sufficient to localize the direction of a real sound source with a hand pointing method while blindfolded (Tabry et al., 2013).

4.2 Small distance changes can be perceived while judging the relative distance of two objects

In the current study, relative distance perception was investigated with a two-alternative forced-choice task. Participants could discriminate which of two virtual objects was closer to them

right down to a difference of 12.9 ± 12.27 cm on average (all distance encoding schemes combined). Performance was better with the encoding scheme which combined intensity modulation with envelope amplitude modulation (10.17 ± 8.42 cm), suggesting that this encoding scheme facilitates relative distance perception within the near space (between 30 and 130 cm). The study by (Richardson et al., 2019), which used a similar paradigm with an SSD and a comparable range of distances, found a distance discrimination score of 8.25 ± 7.26 cm. Although the staircase methods and the size of the targets used in the two studies were not the same, the results are comparable.

Commere and Rouat (2023) compared just noticeable differences (JNDs) in distance between different encoding schemes and found JNDs between 1.6 ± 1.7 cm and 8.8 ± 0.1 cm for stimuli located at a distance of around 95 cm. With the encoding scheme using intensity modulation, Commere and Rouat (2023) found a JND of 6.7 ± 1.8 cm, which is half the discrimination score of 15.63 ± 14.98 cm observed with a similar encoding scheme in the current study. However, the experimental procedure used to determine the JNDs was very different from the discrimination task in the current study since the authors did not used a virtual environment, conducted the experiment online, and used a different protocol. Although the higher performance in Commere and Rouat (2023) was observed with an encoding scheme that used pitch modulation for conveying distance, it has been suggested that this may be an effective and intuitive acoustic feature for conveying elevation information in object recognition (Stiles & Shimojo, 2015) and object localization (Bordeau et al., 2023), even without a familiarization or a training phase.

To accurately estimate the absolute distance of an object with an SSD, it is necessary to be able to perceive differences in the soundscapes depending on the distance at which the object is located. The discrimination score measured in the discrimination task provides us with an insight into the ability to perceive differences in the soundscape depending on the distance of the two objects. We conducted an analysis to test whether participants with higher relative distance localization abilities (indicated by a low discrimination score) also tended to have higher absolute localization abilities (indicated by a low unsigned distance error) and found no correlation between the discrimination score and the unsigned distance error. This finding is consistent with the results of Commere and Rouat, 2023, which suggest that results in a relative distance task (JND in their case) cannot entirely explain accuracy in an absolute distance perception task. One explanation could be that relative and absolute distance perception rely on distinct processes related to internal spatial representation (Kolarik et al., 2016).

5. Conclusion

The current study shows that distance perception with an SSD is compressed, a finding which is in line with the distance compression bias reported in auditory localization experiments. The "floor pointing method" that we developed seems to be an interesting method for assessing absolute distance perception even in the far space. By comparing two distance encoding schemes, both using sound intensity modulation, we showed that a Gaussian-modulation of the sound envelope amplitude can reduce the compression bias and, most importantly, that it reduces the distance overestimation of nearby objects. This result has important practical implications in the context of SSDs developed for locomotion assistance since it could help prevent collisions with obstacles. Although this study still has to be replicated with the actual SSD target population, namely blind people, the experiment was designed in such a way that it can be adapted for use in precisely this population.

Authors Contribution

C.B. and M.A. contributed to the conception and design of the experiment and interpreted the data. C.B. executed the study, performed data analysis and wrote the manuscript in close collaboration with M.A. F.S., C.M. J.D. provided important feedback. All authors have read and approved the manuscript and contributed substantially to it.

Acknowledgements

This research was funded by the Conseil Régional de Bourgogne Franche-Comté (2020_0335), France and the Fond Européen de Développement Régional (FEDER) (BG0027904). The authors thank the Conseil Régional de Bourgogne Franche-Comté, France and the Fond Européen de Développement Régional (FEDER) for their financial support, and the Université de Bourgogne and the Centre National de la Recherche Scientifique (CNRS) for providing administrative and infrastructural support.

6. References

Abboud, S., Hanassy, S., Levy-Tzedek, S., Maidenbaum, S., & Amedi, A. (2014). EyeMusic: Introducing a "visual" colorful experience for the blind using auditory sensory substitution. *Restorative Neurology and Neuroscience, 32*(2), 247–257. https://doi.org/10.3233/RNN-130338

Aladren, A., Lopez-Nicolas, G., Puig, L., & Guerrero, J. J. (2016). Navigation assistance for the visually impaired using RGB-D sensor with range expansion. *IEEE Systems Journal, 10*(3), 922–932. https://doi.org/10.1109/JSYST.2014.2320639

Algazi, V. R., Duda, R. O., Thompson, D. M., & Avendano, C. (2001). The CIPIC HRTF database. *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics,* 99–102. https://doi.org/10.1109/ASPAA.2001.969552

Ambard, M., Benezeth, Y., and Pfister, P. (2015). Mobile video-to-audio transducer and motion detection for sensory substitution. *Frontiers in ICT 2*, 20. https://doi.org/10.3389/fict.2015.00020

Auvray, M. (2004). *Immersion et perception spatiale. L'exemple des dispositifs de substitution sensorielle*. Ecole des Hautes Etudes en Sciences Sociales, Paris.

Auvray, M., Hanneton, S., & O'Regan, J. K. (2007). Learning to perceive with a visuo-auditory substitution system: Localisation and object recognition with 'The Voice'. *Perception, 36*(3), 416–430. https://doi.org/10.1068/p5631

Bazilinskyy, P., van Haarlem, W., Quraishi, H., Berssenbrugge, C., Binda, J., & de Winter, J. (2016). Sonifying the location of an object: A comparison of three methods. *IFAC-PapersOnLine*, 49(19), 531–536. https://doi.org/10.1016/j.ifacol.2016.10.614

Blauert, J. (1983). *Spatial hearing : The psychophysics of human sound localization*. MIT Pr.

Bordeau, C., Scalvini, F., Migniot, C., Dubois, J., & Ambard, M. (2023). Cross-modal correspondence enhances elevation localization in visual-to-auditory sensory substitution. *Frontiers in Psychology, 14*. https://doi.org/10.3389/fpsyg.2023.1079998

Bronkhorst, A. W., & Houtgast, T. (1999). Auditory distance perception in rooms. *Nature, 397*(6719), 517–520. https://doi.org/10.1038/17374

Capelle, C., Trullemans, C., Arno, P., & Veraart, C. (1998). A real-time experimental prototype for enhancement of vision rehabilitation using auditory substitution. *IEEE Transactions on Biomedical Engineering, 45*(10), 1279–1293. https://doi.org/10.1109/10.720206

Commère, L., & Rouat, J. (2023). Evaluation of short range depth sonifications for visual-to-auditory sensory substitution (arXiv:2304.05462). arXiv. http://arxiv.org/abs/2304.05462

Cronly-Dillon, J., Persaud, K., & Gregory, R. P. F. (1999). The perception of visual images encoded in musical form: A study in cross-modality information transfer. *Proceedings of the Royal Society of London. Series B: Biological Sciences, 266*(1436), 2427–2433. https://doi.org/10.1098/rspb.1999.0942

Grassi, M., & Casco, C. (2009). Audiovisual bounce-inducing effect: Attention alone does not explain why the discs are bouncing. *Journal of Experimental Psychology: Human Perception and Performance, 35*(1), 235–243. https://doi.org/10.1037/a0013031

Hamilton-Fletcher, G., Alvarez, J., Obrist, M., & Ward, J. (2022). SoundSight: A mobile sensory substitution device that sonifies colour, distance, and temperature. *Journal on Multimodal User Interfaces, 16*(1), 107–123. https://doi.org/10.1007/s12193-021-00376-w

Hamilton-Fletcher, G., Mengucci, M. and Medeiros, F. (2016). Synaestheatre: sonification of coloured objects in space, in: *Proceedings of the 2016 International Conference on Live Interfaces*, pp. 252–256. Brighton, UK. https://doi.org/10.13140/RG.2.1.5053.7845

Hamilton-Fletcher, G., Obrist, M., Watten, P., Mengucci, M., & Ward, J. (2016). « I always wanted to see the night sky » : Blind user preferences for sensory substitution devices. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2162–2174. https://doi.org/10.1145/2858036.2858241

Hanneton, S., Auvray, M., & Durette, B. (2010). The Vibe : A versatile vision-to-audition sensory substitution device. *Applied Bionics and Biomechanics, 7*(4), 269–276. https://doi.org/10.1080/11762322.2010.512734

Kayukawa, S., Higuchi, K., Guerreiro, J., Morishima, S., Sato, Y., Kitani, K., & Asakawa, C. (2019). BBeep: A sonic collision avoidance system for blind travellers and nearby pedestrians. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3290605.3300282

Kolarik, A. J., Cirstea, S., & Pardhan, S. (2013). Evidence for enhanced discrimination of virtual auditory distance among blind listeners using level and direct-to-reverberant cues. *Experimental Brain Research, 224*(4), 623–633. https://doi.org/10.1007/s00221-012-3340-0

Kolarik, A. J., Moore, B. C. J., Zahorik, P., Cirstea, S., & Pardhan, S. (2016). Auditory distance perception in humans: A review of cues, development, neuronal bases, and effects of

sensory loss. *Attention, Perception, & Psychophysics, 78*(2), 373–395. https://doi.org/10.3758/s13414-015-1015-1

Kolarik, A. J., Pardhan, S., Cirstea, S., & Moore, B. C. J. (2017). Auditory spatial representations of the world are compressed in blind humans. *Experimental Brain Research, 235*(2), 597–606. https://doi.org/10.1007/s00221-016-4823-1

Kolarik, A. J., Raman, R., Moore, B. C. J., Cirstea, S., Gopalakrishnan, S., & Pardhan, S. (2020). The accuracy of auditory spatial judgments in the visually impaired is dependent on sound source distance. *Scientific Reports, 10*(1). https://doi.org/10.1038/s41598-020-64306-8

Kopčo, N., & Shinn-Cunningham, B. (2011). Effect of stimulus spectrum on distance perception for nearby sources. *Journal of the Acoustical Society of America, 130*(3), 1530–1541. https://doi.org/10.1121/1.3613705

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package : Tests in Linear Mixed Effects Models. *Journal of Statistical Software, 82*(13). https://doi.org/10.18637/jss.v082.i13

Lenth, R. V. (2022). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.7.4–1

Macé, M. J.-M., Dramas, F., & Jouffrais, C. (2012). Reaching to Sound Accuracy in the Peri-personal Space of Blind and Sighted Humans. In *Lecture Notes in Computer Science* (p. 636–643). https://doi.org/10.1007/978-3-642-31534-3_93

Begault, D. R. (1995). 3-D sound for virtual reality and multimedia. *Computer Music Journal, 19*(4), 99. https://doi.org/10.2307/3680997

Martin, V., Viaud-Delmon, I., & Warusfel, O. (2021). Effect of Environment-Related Cues on Auditory Distance Perception in the Context of Audio-Only Augmented Reality. *Applied Sciences, 12*(1), 348. https://doi.org/10.3390/app12010348

Meijer, P. B. L. (1992). An experimental system for auditory image representations. *IEEE Transactions on Biomedical Engineering, 39*(2), 112–121. https://doi.org/10.1109/10.121642 https://doi.org/10.1109/10.121642

Mhaish, A., Gholamalizadeh, T., Ince, G., & Duff, D. J. (2016). Assessment of a visual to spatial-audio sensory substitution system. *2016 24th Signal Processing and Communication Application Conference (SIU)*, 245–248. https://doi.org/10.1109/SIU.2016.7495723 https://doi.org/10.1109/SIU.2016.7495723

Mohlin, P. (2011). The just audible tonality of short exponential and Gaussian pure tone bursts. *The Journal of the Acoustical Society of America, 129*(6), 3827–3836. https://doi.org/10.1121/1.3573990 https://doi.org/10.1121/1.3573990

Negen, J., Bird, L.-A., Slater, H., Thaler, L., & Nardini, M. (2023). Multisensory perception and decision-making with a new sensory skill. *Journal of Experimental Psychology: Human Perception and Performance, 49*(5), 600–622. https://doi.org/10.1037/xhp0001114 https://doi.org/10.1037/xhp0001114

Negen, J., Wen, L., Thaler, L., & Nardini, M. (2018). Bayes-like integration of a new sensory skill with vision. *Scientific Reports, 8*(1). https://doi.org/10.1038/s41598-018-35046-7

Neugebauer, A., Rifai, K., Getzlaff, M., & Wahl, S. (2020). Navigation aid for blind persons by visual-to-auditory sensory substitution: A pilot study. *PLOS ONE, 15*(8), e0237344. https://doi.org/10.1371/journal.pone.0237344

Neuhoff, J. G. (1998). Perceptual bias for rising tones. *Nature, 395*(6698), 123–124. https://doi.org/10.1038/25862

Parseihian, G., Jouffrais, C., & Katz, B. F. G. (2014). Reaching nearby sources: Comparison between real and virtual sound and visual targets. *Frontiers in Neuroscience, 8*. https://doi.org/10.3389/fnins.2014.00269

Parseihian, G., Katz, B. F., & Conan, S. (2012). Sound effect metaphors for near field distance sonification. *2012 International Conference on Auditory Display (ICAD)*

Plack, C. J., & Oxenham, A. J. (2006). *The psychophysics of pitch.* Dans *Springer eBooks* (p. 7–55). https://doi.org/10.1007/0-387-28958-5_2

Renier, L., & De Volder, A. G. (2010). Vision substitution and depth perception: Early blind subjects experience visual perspective through their ears. *Disability and Rehabilitation: Assistive Technology, 5*(3), 175–183. https://doi.org/10.3109/17483100903253936

Ribeiro, F., Florencio, D., Chou, P. A., & Zhang, Z. (2012). Auditory augmented reality: Object sonification for the visually impaired. *2012 IEEE 14th International Workshop on Multimedia Signal Processing (MMSP)*, 319–324. https://doi.org/10.1109/MMSP.2012.6343462

Richardson, M., Thar, J., Alvarez, J., Borchers, J., Ward, J., & Hamilton-Fletcher, G. (2019). How much spatial information is lost in the sensory substitution process? Comparing visual, tactile, and auditory approaches. *Perception, 48*(11), 1079–1103. https://doi.org/10.1177/0301006619873194

Ries, D. T., Schlauch, R. S., & DiGiovanni, J. J. (2008). The role of temporal-masking patterns in the determination of subjective duration and loudness for ramped and damped sounds. *The Journal of the Acoustical Society of America, 124*(6), 3772–3783. https://doi.org/10.1121/1.2999342

Rossing, T. D., & Houtsma, A. J. M. (1986). Effects of signal envelope on the pitch of short sinusoidal tones. *The Journal of the Acoustical Society of America, 79*(6), 1926–1933. https://doi.org/10.1121/1.393199

Schutz, M., & Gillard, J. (2020). On the generalization of tones: A detailed exploration of non-speech auditory perception stimuli. *Scientific Reports, 10*(1). https://doi.org/10.1038/s41598-020-63132-2

Sekuler, R., Sekuler, A. B., & Lau, R. (1997). Sound alters visual motion perception. *Nature, 385*(6614), 308–308. https://doi.org/10.1038/385308a0

Shinn-Cunningham, B. G. (2000). Distance cues for virtual auditory space. *Proceedings of the IEEE Pacific Rim Conference (PRC) on Multimedia*.

Sodnik, J., Sušnik, R., Štular, M., & Tomažič, S. (2005). Spatial sound resolution of an interpolated HRIR library. *Applied Acoustics, 66*(11), 1219–1234. https://doi.org/10.1016/j.apacoust.2005.04.003

Stiles, N. R. B., & Shimojo, S. (2015). Auditory sensory substitution is intuitive and automatic with texture stimuli. *Scientific Reports, 5*(1). https://doi.org/10.1038/srep15628

Stoll, C., Palluel-Germain, R., Fristot, V., Pellerin, D., Alleysson, D., & Graff, C. (2015). Navigating from a depth image converted into sound. *Applied Bionics and Biomechanics*, 1–9. https://doi.org/10.1155/2015/543492

Tabry, V., Zatorre, R. J., & Voss, P. (2013). The influence of vision on sound localization abilities in both the horizontal and vertical planes. *Frontiers in Psychology, 4*. https://doi.org/10.3389/fpsyg.2013.00932

Team, R. C. (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R Core Team.

Vallet, G. T., Shore, D. I., & Schutz, M. (2014). Exploring the role of the amplitude envelope in duration estimation. *Perception, 43*(7), 616–630. https://doi.org/10.1068/p7656

Voss, P., Lassonde, M., Gougoux, F., Fortin, M., Guillemot, J.-P., & Lepore, F. (2004). Early- and late-onset blind individuals show supra-normal auditory abilities in far-space. *Current Biology, 14*(19), 1734–1738. https://doi.org/10.1016/j.cub.2004.09.051

Yost, W. A. (2017). Sound source localization identification accuracy: Envelope dependencies. *The Journal of the Acoustical Society of America, 142*(1), 173–185. https://doi.org/10.1121/1.4990656

Zahorik, P. (2002). Assessing auditory distance perception using virtual acoustics. *The Journal of the Acoustical Society of America, 111*(4), 1832–1846. https://doi.org/10.1121/1.1458027

Zahorik, P. (2005). Auditory distance perception in humans: A summary of past and present research. *Acta Acoustica United with Acustica, 91*, 409–420.

2.3. Synthèse

L'Étude 2 avait pour objectif d'évaluer les capacités de localisation de la distance d'objets avec le DSS en comparant deux schémas d'encodage pour la distance (tous les deux basés sur le schéma d'encodage *Monotonic* de l'Étude 1 pour les dimensions de l'azimut et de l'élévation). Les deux schémas d'encodage pour la distance reposaient sur la modulation de l'intensité en se rapprochant de l'atténuation de l'intensité observée lors de la propagation de sources sonores réelles, mais l'un intégrait une modulation gaussienne de l'enveloppe, modifiant le timbre.

En proposant un nouveau protocole d'évaluation des capacités de perception de la distance d'un objet avec un DSS (pointage au sol), les résultats ont confirmé la présence d'un biais de compression dans la perception de la distance avec les deux schémas d'encodage, comme ce qui est couramment observé dans la perception spatiale auditive. Ces résultats confirment que l'utilisation d'indices acoustiques spatiaux (atténuation de l'intensité) dans le schéma d'encodage d'un DSS fait face à des limites perceptives similaires aux limites de la perception spatiale auditive.

De plus, les résultats ont montré que cette perception compressée avec le DSS pouvait être réduite, notamment avec une modulation gaussienne de l'enveloppe sonore. Cette meilleure estimation de la distance peut être attribuée à la modification du timbre des sons composant le paysage sonore, ou bien à une modulation différente de la sonie avec ce schéma d'encodage. Les résultats de la tâche de discrimination suggèrent également de meilleures capacités avec le schéma d'encodage utilisant la modulation gaussienne de l'enveloppe, mais pour discriminer la distance de deux objets.

Dans le contexte du développement de DSS vision-vers-audition pour l'aide à la locomotion et à la localisation d'obstacles, il est important de réduire la surestimation des distances proches pour réduire la probabilité de collision avec les obstacles. Les résultats de l'Étude 2 suggèrent qu'aux premiers stades de l'utilisation du DSS, la perception de la distance est bien compressée mais qu'il est possible de compenser en partie ce biais. Enfin, cette étude a permis de valider un schéma d'encodage 3-dimensionnelle en évaluant les capacités de localisation avec un nouveau protocole dans des conditions d'usage d'un DSS plus proches de la réalité, en laissant la possibilité aux participants de bouger la tête.

3. Étude 3

Les capacités de localisation avec un dispositif de substitution vision-vers-audition sont modulées par la configuration spatiale de la scène

Localization abilities with a visual-to-auditory substitution device are modulated by the spatial arrangement of the scene

Camille Bordeau, Florian Scalvini, Cyrille Migniot, Julien Dubois and Maxime Ambard

À soumettre

3.1. Résumé

Les dispositifs de substitution sensorielle vision-vers-audition convertissent des images en paysages sonores. Ils sont destinés à être utilisés par les personnes non-voyantes lors de leurs déplacements pédestres au quotidien durant lesquels plusieurs obstacles doivent être localisés simultanément, et d'autres objets non-pertinents doivent être ignorés. Il est donc primordial d'établir dans quelle mesure les dispositifs de substitution permettent de localiser des obstacles dans des scènes complexes.

Dans cette étude, nous utilisons un dispositif de substitution qui transmet des informations spatiales en combinant des indices acoustiques spatiaux et une modulation de hauteur. Nous avons évalué la capacité à effectuer une tâche de localisation dans une scène virtuelle minimaliste ou complexe avec 19 participants voyants aux yeux bandés qui devaient pointer une cible virtuelle présentée seule ou parmi des distracteurs. La configuration spatiale de la scène a été manipulée en faisant varier le nombre de distracteurs (0, 2 ou 4) et leur disposition spatiale relativement à la cible (alignés horizontalement, verticalement ou non-alignés).

Alors que les capacités de localisation pour la dimension de l'élévation n'étaient pas altérées par la présence des distracteurs, les performances de localisation en azimut étaient modulées lorsque de nombreux distracteurs étaient affichés à la même élévation que la cible (i.e., alignés horizontalement).

Les performances de localisation de l'élévation tendent à confirmer l'efficacité de la modulation de la hauteur tonale pour transmettre des informations sur l'élévation avec un dispositif de substitution vision-vers-audition en préservant les capacités de ségrégation dans diverses configurations spatiales. Inversement, les difficultés de localisation de l'azimut semblent résulter de difficultés de ségrégation survenant lorsque la configuration spatiale des objets ne permet pas une ségrégation aisée de la hauteur. L'ensemble de ces résultats suggère qu'aux premiers stades de l'utilisation d'un dispositif de substitution, il peut être difficile de ségréguer les obstacles dans la scène lorsqu'ils sont alignés horizontalement. Cet aspect doit être pris en compte dans la conception des dispositifs de substitution afin d'aider les personnes non-voyantes à appréhender correctement la dangerosité des situations.

3.2. Article

Localization abilities with a visual-to-auditory substitution device are modulated by the spatial arrangement of the scene

Camille Bordeau[1], Florian Scalvini[2], Cyrille Migniot[2], Julien Dubois[2] and Maxime Ambard[1]

[1] LEAD-CNRS UMR5022, Université de Bourgogne, Dijon, France
[2] ImViA EA 7535, Université de Bourgogne, Dijon, France

Abstract

Introduction: Visual-to-auditory sensory substitution devices convert visual images into soundscapes. They are meant to be used by blind people during walking trips in daily situations with several obstacles that must be localized simultaneously, but also irrelevant objects that must be ignored. It is thus of first importance to establish to what extent substitution devices allow the localization of obstacles among complex scenes.

Method: In this study, we use a substitution device that transmits spatial information by combining spatial acoustic cues and pitch modulation. We evaluate the ability to perform a localization task in minimalist and complex virtual scenes with 19 blindfolded sighted participants that had to point to a virtual target that was displayed alone or among distractors. The spatial configuration of the scene was manipulated by varying the number of distractors (0, 2 or 4) and their spatial disposition regarding the target (horizontally aligned, vertically aligned, or not aligned).

Results: While elevation localization abilities are not impaired by the presence of distractors, performance to localize the azimuth of the target is modulated when many distractors are displayed at the same elevation as the target (horizontally aligned).

Discussion: Performance to localize the elevation tends to confirm the effectiveness of the pitch modulation to convey elevation information with a visual-to-auditory substitution device by preserving segregation abilities in various spatial configurations. Conversely, the localization impairment for the azimuth seems to result from segregation difficulties occurring when the spatial configuration of the objects does not allow pitch segregation. Taken together, these results suggest that, at the early stage of use of a substitution device, it can be difficult to spatially segregate obstacles when they are horizontally aligned. This must be considered in the design of substitution devices in order to help blind people correctly evaluate the dangerousness of situations.

Keywords: Sensory substitution, cocktail party, sonification, image-to-sound conversion, localization, visual impairment, auditory scene analysis, feature segregation

1. Introduction

Sensory substitution devices (SSDs) aim at transmitting information about the surrounding environment through a remaining sensory modality. Visual-to-auditory SSDs convert visual information into soundscapes by mapping visual features into auditory cues. Despite promising results on the feasibility of visual-to-auditory SSDs, they are still not well adopted by the blind in daily life.

It is commonly stipulated that the auditory information provided by a visual-to-auditory SSD can result in an auditory overload when environmental sounds are heard or when the visual scene perceived with the SSD is complex (Elli et al., 2014; Maidenbaum et al., 2014). In daily life, the surrounding environment comprises multiple objects, which may cause difficulties in interpreting the auditory information that would be provided by the SSD. For this reason, the perceptual segmentation of the soundscape is an important capacity to evaluate in order to improve the usability of SSDs to localize obstacles, as stated in Hamilton-Fletcher & Chan (2021). However, most of the studies on SSD use simple scene configurations. Studies with localization tasks usually display a unique object in the scene (Ambard et al., 2015; Auvray et al., 2007; Bordeau et al., 2023; Brown et al., 2011; Commere et al., 2020; Hanneton et al., 2010; Levy-Tzedek et al., 2012; Mhaish et al., 2016; Pourghaemi et al., 2018; Proulx et al., 2008) although it is rarely the case outside a laboratory context. In a real context of use, the relevant SSD information has to be separated from natural sounds (e.g., people talking, car klaxons) as well as SSD information that might not be of high importance for pedestrian mobility.

The work of Buchs et al. (2019) investigated the effect of irrelevant background sounds on the ability to perform a task with the EyeMusic SSD (Abboud et al., 2014). Encouragingly, they showed the ability of blind participants to efficiently use the SSD soundscapes for identifying the color and shape of visual stimuli, while irrelevant environmental sounds had to be ignored. Therefore, similarly to what is reported in the cocktail party problem (Bronkhorst, 2000; Cherry, 1953), participants in the study of Buchs et al. (2019) were able to focus their attention on the SSD soundscapes played through bone-conduction headphones while ignoring the irrelevant background noise. However, in this study, the irrelevant auditory information was played from a real sound source, and they therefore did not directly evaluate the ability to segregate relevant information within complex synthesized SSD soundscapes.

Some studies have investigated the ability to use an SSD while distinct objects were present in the scene and transmitted through the SSD soundscapes. For instance, Richardson et al. (2019) showed the feasibility of segregating two objects perceived with the SSD Synaestheatre while they were located at distinct distances or distinct elevation locations. To a certain extent, participants in

their study were able to discriminate the distance, or the elevation, of the two objects if they were sufficiently spatially separated. In Ambard et al. (2015), they observed difficulties in segregating two objects perceived with an SSD when they were simultaneously displayed at the same elevation. In their work, Brown et al. (2015) investigated with the vOICe SSD (Meijer, 1992) the ability to segregate two distinct lines that were sonified with the SSD, depending on the consonance (frequency component) of the resulting SSD soundscape. They found that the perceptive segregation of the two horizontal lines into distinct objects was impaired when the SSD soundscape contained consonant harmonic relations. However, in those previous SSD studies, the two displayed objects (or visual features) were relevant for the tasks, while in a real context of SSD use, many irrelevant objects may be present in the environment and have to be ignored to process the relevant information.

The abilities to localize a real sound source played among an irrelevant acoustic background have been assessed in previous studies using sound maskers (Brungart et al., 2005; Brungart et al., 2014; Lorenzi et al., 1999). In these cocktail party configurations, sequential localization tasks were used during which a broadband noise (Brungart et al., 2005) or a broadband environmental sound source to localize (Brungart et al., 2014) was "added" among up to 5 irrelevant sounds in Brungart et al. (2014) and 13 irrelevant sounds in Brungart et al. (2005). These studies showed a decrease in localization performance with an increasing number of concurrent sound sources, which was severe when more than 5 sound sources were played simultaneously. These studies using real sound sources give insights into the limits of separating relevant sound sources in an ecological context where the participants can use natural auditory cues.

The capacities of separating sound sources from a background have also been investigated with simulated sound sources (Best et al., 2004; Feierabend et al., 2019; Kawashima & Sato, 2015). For instance, using sound sources spatialized with individualized HRTFs, Best et al. (2004) investigated the ability to spatially segregate two broadband sound sources simultaneously played while being separated either in azimuth or in elevation. They showed that the abilities to segregate the two sound sources depended on their spatial alignment (azimuth or elevation). When the sound sources were aligned along the same elevation but located at different azimuths around the median axis, the azimuth separation required to perceive two distinct sound sources was lower than when the two sound sources were located more laterally. In contrast, when the sound sources were aligned along the same azimuth but located at different elevations, segregation abilities were lower when the sound sources were located closer to the median axis than laterally. In another study with simulated environmental sounds presented alone or with four other sound sources, Feierabend et

al. (2019) showed a decrease in localization performance with both sighted blindfolded and blind participants in the cocktail party configuration.

These studies show clear evidence that localizing a simulated sound source among other sounds that are simultaneously played may result in localization impairments. Furthermore, the spatial arrangement of the sound sources influences this effect (Kwak & Han, 2020). For instance, the spatial separation between the sound sources reduces the localization impairments (Kawashima & Sato, 2015) which is an effect known as the spatial release from masking, and the segregation abilities are influenced by the dimension (azimuth or elevation) along which the sound sources are aligned (Best et al., 2004).

It is thus well established that, in the context of auditory scene analysis, the abilities to separate (and localize) a real or simulated sound source from an irrelevant background are limited and depend both on the spatial arrangement of the auditory scene and on the number of the sound sources. However, in the context of SSDs that are intended for use in complex situations with multiple simultaneous obstacles, this has never been directly investigated. For this reason, this study evaluates the ability to use an SSD to localize an object in complex scenes where multiple irrelevant objects are also displayed and transmitted in the SSD soundscape, considering both the number of objects and their spatial arrangement.

Perception abilities were assessed with a pointing localization task after a brief familiarization with the SSD encoding scheme. The SSD encoding scheme uses spatial binaural acoustic cues for the azimuth dimension and combines spatial spectral acoustic and pitch modulation for the elevation dimension. Since the number of simultaneous sound sources and their spatial disposition are known to influence localization performance, the number of simultaneous distractors in the scene was manipulated (0, 2 and 4) as well as their spatial disposition relative to the target (either horizontally aligned, vertically aligned or non-aligned). We predicted a decrease in localization performance with increasing scene complexity (i.e., with an increasing number of distractors). We expected a more pronounced effect of the scene complexity on azimuth localization abilities than elevation localization abilities, since the pitch modulation for the elevation dimension should result in relevant spectral information. Also, we hypothesized that localization abilities should be less impaired when the spatial disposition of the objects in the scene results in object-specific spectral signatures in the SSD soundscape (i.e., when the objects are located at distinct elevations).

2. Method

2.1 Participants

The study included 19 participants (age: M = 23.7, SD = 3.3, 14 males, 17 right-handed). None of the participants reported any hearing impairments, psychiatric illnesses, or neurological disorders in their medical histories. The experimental protocol received approval from the local ethical committee Comité d'Ethique pour la Recherche de Université Bourgogne Franche-Comté (CERUBFC-2021-12-21-050) and followed the ethical guidelines of the Declaration of Helsinki. All participants provided written informed consent before participating in the experiment and did not receive any monetary compensation.

2.2 Material and apparatus

2.2.1 Virtual environment

The experiment was conducted in a minimalist UNITY3D virtual environment composed of a virtual camera and virtual objects (the target and the distractors). The virtual camera and the pointing tool were respectively associated with the participants' heads and with a gun pointing tool that were tracked using HTC VIVE Trackers 2.0 and monitored with 4 HTC VIVE base stations. The virtual environment could not be visually explored since the participants were blindfolded and did not wear the virtual reality headset.

2.2.2 Visual-to-auditory SSD

The visual-to-auditory SSD converts in real time a video stream into soundscapes containing the 3-dimensional spatial information, as explained in the following sections.

2.2.2.1 Video acquisition and processing

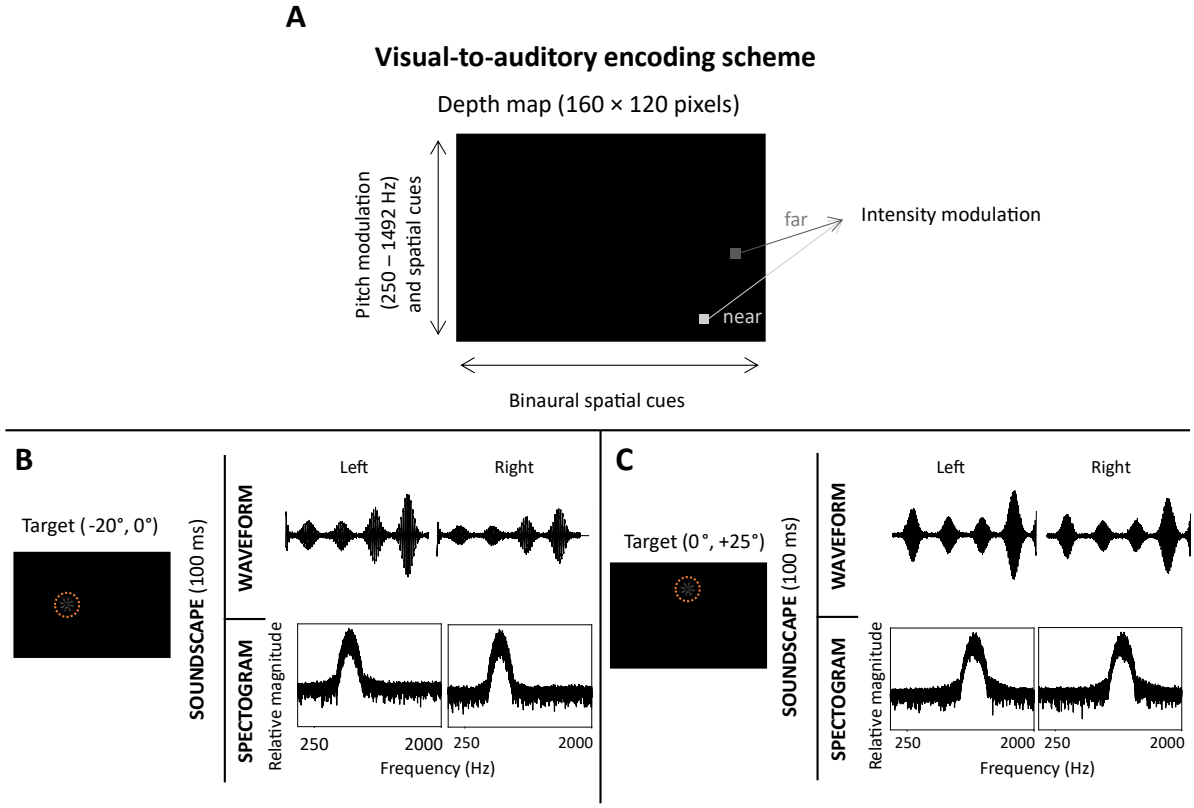
Video is acquired with a virtual camera with a field of view of 90 × 74° (Horizontal × Vertical), and a frame rate of 60 Hz. The raw video consists of a depth map encoded into grayscale images ranging from black (5.01 m) to white (0.01 m) gray levels. The processed video frame contains pixels for which the absolute difference in gray levels between consecutive frames (frame differencing) is larger than a threshold of 10. The processed grayscale image is then scaled into a dimension of 160 × 120 pixels (Horizontal × Vertical) which contains only new visual information, i.e., the 'active' graphical pixels selected by the video processing.

2.2.2.2 Visual-to-auditory conversion

The extracted visual features (i.e., the active pixels) contained in the processed video frame are then converted into a soundscape following an encoding scheme mapping 3-dimensional

spatial information (azimuth, elevation, and distance) with acoustic cues. The visual-to-auditory encoding scheme is schematized in Figure 1. The soundscape is composed of summed 'auditory pixels' constituting the audio frame, and consecutive audio frames are combined in real time to form a continuous audio stream. An auditory pixel is a 35 ms stereophonic spatialized Gaussian-modulated monotone that is associated with both the location and the gray level of a given graphical pixel position in the processed image. Elevation location is mapped to the pitch of the monotone (from 250 Hz to 1492 Hz following the Mel scale), so the lower pitches correspond to lower elevation locations. The monotone is spatialized in azimuth and elevation using HRTFs from the CIPIC database (Algazi et al., 2001). For the distance dimension, the sound intensity of the auditory pixel is modulated such that the sound intensity increases with increasing gray level (i.e., decreasing distance). Soundscapes were delivered in real time with a SONY MDR-7506 headphone.

Figure 1. Visual-to-auditory encoding scheme (**A**). Azimuth is conveyed with binaural cues, elevation with pitch modulation and spatial cues, and distance is conveyed with intensity modulation. Two examples of 100 ms soundscapes (bottom) corresponding to a target (surrounded in orange) localized on the left side (**B**) and upper part of the processed depth-map image (**C**). The spectrum shows the difference in frequency components between the two locations (higher frequencies for the upper position). The waveform of the left location shows the binaural spatial cues.

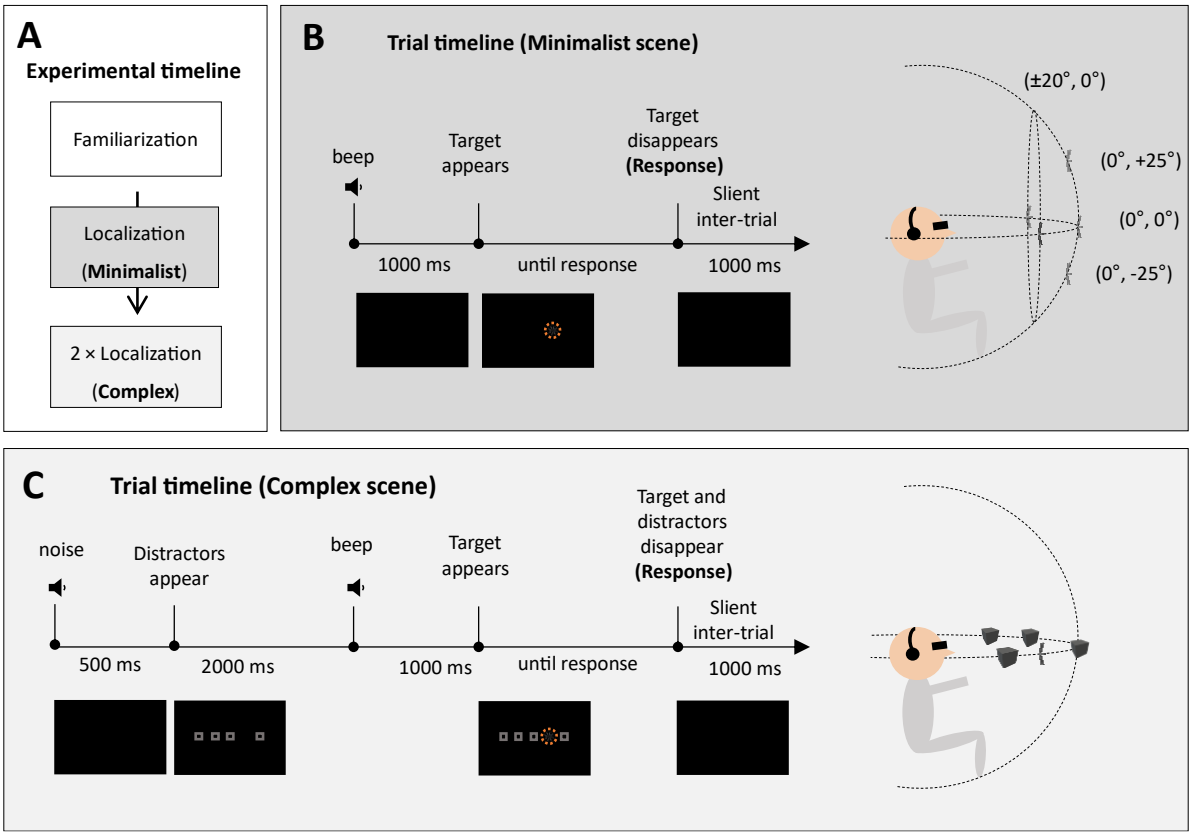


2.2.3 Virtual objects

2.2.3.1 Virtual target

The virtual target to localize was a self-rotating 3D propeller shape consisting of 4 intersecting bars of a length of 5 cm, identical to the one used in Bordeau et al. (2023). The virtual target was systematically placed on a 1 meter-radius virtual sphere centered on the position of the virtual camera (Figure 2) at one of the five following positions: centered in azimuth at a high position (azimuth = 0°, elevation = +25°), middle position (0°, 0°), or bottom position (0°, -25°), or laterally on the right side (+20°, 0°) or left side (-20°, 0°).

Figure 2. Experimental timeline (**A**) and trial timeline in the Minimalist scene condition (**B**) and Complex scene condition (**C**). (**A**) Right after the brief verbal explanations of the SSD encoding scheme and the short audio-motor familiarization, participants practiced a block of localization task in the Minimalist scene (without distractors), followed by two blocks in the Complex scene (with distractors). (**B**) In the Minimalist scene trials, the target (the propeller surrounded in orange) appeared after a short beep auditory signal and disappeared when the participants pressed the pointing tool to record the location of its response. The 5 possible target locations are depicted in the right figure (opaque and transparent gray propeller). (**C**) In the Complex scene trials, the apparition of the target (propeller surrounded in orange) was preceded by the apparition of the distractors (gray cubes), which remained displayed until the participants pressed the pointing tool. The experimental view is schematized on the right figure, in which the target (opaque propeller) is located at the right location (+20°, 0°) with 4 horizontally aligned distractors (gray cubes).



2.2.3.2 Virtual distractors

The virtual distractors were self-rotating cubes of a dimension of 8 x 8 x 8 cm displayed on a 2-meter-radius sphere centered on the virtual camera. The number of distractors (Number: 2, or 4) and their spatial disposition relative to the target (Disposition: Non-aligned, Vertically-aligned, or Horizontally-aligned) were manipulated in a within-subject design. In the Vertically-aligned condition, the distractors and the target were vertically aligned on the image, ranging from -25° to +25° in elevation, equally spaced by 25° with 2 distractors and by 12.5° with 4 distractors. In the Horizontally-aligned condition, distractors were horizontally aligned with the target, separated by 20° in azimuth along the azimuth range from -20° to +20° with 2 distractors and from -40° to +40° with 4 distractors. In the Vertically-aligned and Horizontally-aligned conditions, no distractor was displayed at the location of the target, so the target was always separated from the nearest distractor at least by 12.5° in elevation and 20° in azimuth. In the Non-aligned condition, the distractors were never aligned with the target since their coordinates were (-30°, +27°) and (+30°, -27°) with 2 distractors, and 2 additional positions (-30°, -27°) and (+30°, +27°) with 4 distractors. The three spatial dispositions of the distractors relative to the target are depicted in Figure 3.

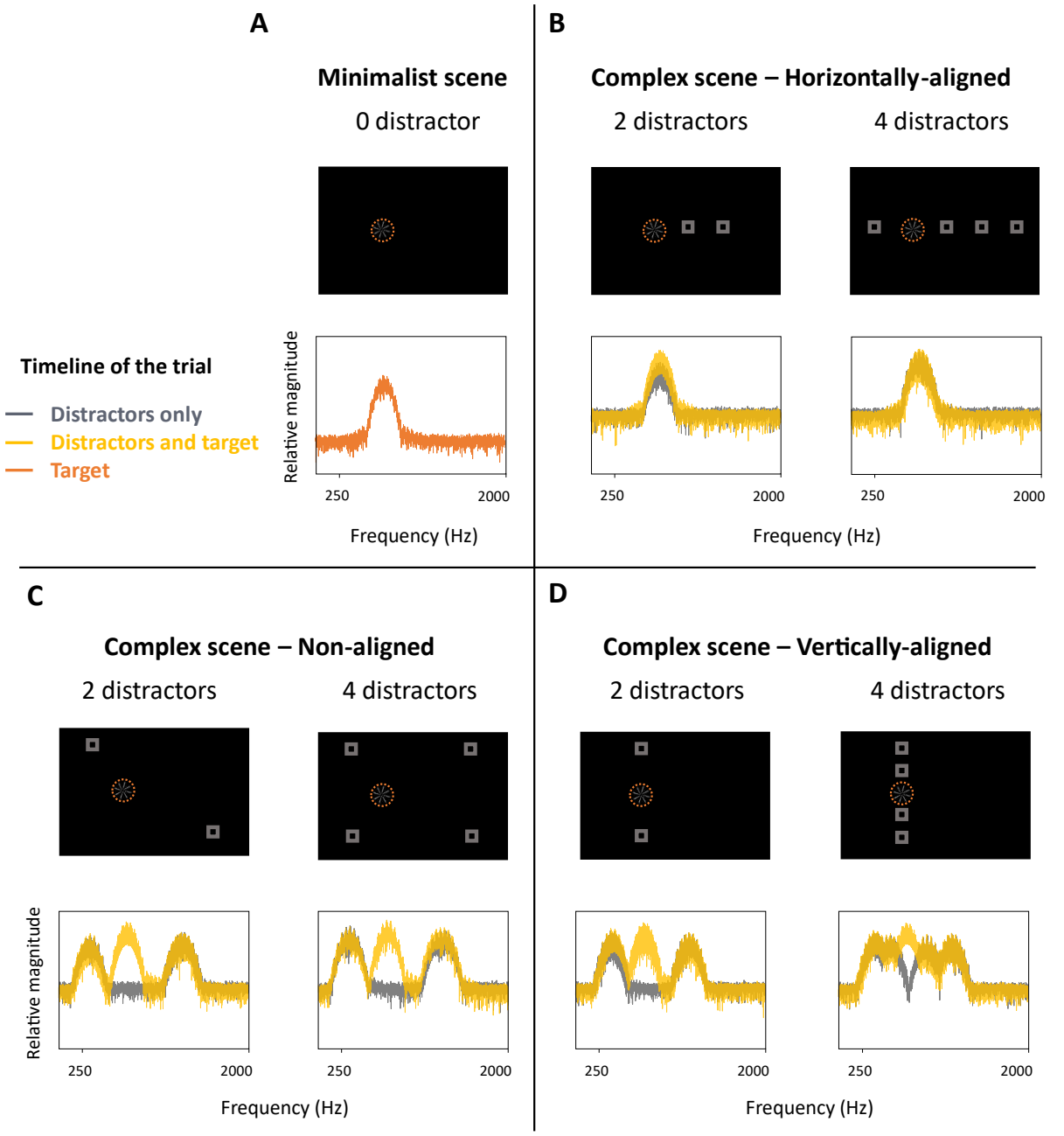
2.3 Experimental procedure

The experiment consisted of a unique session of 45 minutes. After participants gave their informed consent and filled out a demographic questionnaire, the experimenter briefly explained the main principles of the SSD. Participants then followed an active familiarization with the SSD and practiced the 3 blocks of the localization task, each composed of 48 trials. The timeline of a session is given in Figure 2 and explained more in more detail in the following sections.

2.3.1 Verbal explanations of the main principles of the SSD

The experimenter gave verbal explanations to the participant about the conversion principles of the SSD at the beginning of the experiment. Participants were informed that they would have to localize a virtual target while being blindfolded, based on sounds that depend on the lateral and vertical position of the target in their front space. The encoding scheme for azimuth (spatialization) and elevation (pitch modulation) was briefly explained to the participants. Since the distance was not manipulated during the experiment, no information was transmitted to the participants about this dimension.

Figure 3. Processed image captured by the virtual camera and frequency spectrum during a trial as a function of the Scene (Minimalist **(A)** or Complex **(B)**, **(C)** and **(D)**, the Number of distractors (0, 2 or 4), and their Spatial disposition (Horizontally-aligned, Non-aligned or Vertically-aligned). The target (surrounded in orange) was located at the left location (-20°, 0°). The frequency spectrum of the left ear channel corresponding to the SSD soundscapes associated with the processed image is provided below, depending on the timeline of the trial. In the Minimalist scene (without distractor), the frequency spectrum corresponds to the soundscape when the target is displayed (Target, orange). In the Complex scene (with distractors), the frequency spectrum is provided separately for the phase during which only the distractors are displayed (Distractors only, gray), and for the phase during which the target is displayed among the distractors (Distractors and target, yellow).



2.3.2 Audio-motor familiarization with the SSD

Immediately after the verbal explanations, participants started an active familiarization for 90 seconds while being blindfolded. They were instructed to direct the pointing tool to their front space with the arm stretched. The virtual target was continuously placed on a 1-meter-radius virtual sphere centered on the position of the virtual camera, at the intersection between the ray coming from the pointing tool and the sphere. Participants were instructed to pay attention to the sounds they heard, depending on the location of the virtual target. They were free to place the target wherever they wanted, but they were encouraged to place the target at various elevations and azimuth positions and to pay attention to the spatial limit of the space where the target could be heard (i.e., the field of sonification). Since the position of the virtual camera was updated only at the beginning of the familiarization with the participant's head tracker position, they were instructed to keep their heads still. The participant's head tracker position was recorded during the familiarization session to check that they followed this instruction.

2.3.3 Localization task

After the familiarization, participants practiced 3 blocks of 48 trials of the localization task. For all participants, the first block was the localization task without distractors (Minimalist scene), while the second and third blocks were with distractors (Complex scene). In the three blocks, blindfolded participants had to localize the virtual target by pointing to it with the pointing tool, based on soundscapes provided by the SSD. They were asked to point to the target as accurately as possible. The position of the virtual camera was updated at the beginning of each trial with the participant's head tracker position. Participants were blindfolded during all blocks and were instructed to keep their heads still. The head tracker position was recorded during the localization task to check that they followed this instruction.

2.3.3.1 Minimalist scene (without distractor)

In the Minimalist scene, the target was displayed without a distractor (Figure 2, top panel). Each of the 48 trials began with a 500 ms auditory signal (a 400-hz beep) indicating that the virtual target was going to be displayed in 500 ms. The virtual target was then displayed at one of the 5 possible locations until the participant pressed the trigger to log the perceived position.

2.3.3.2 Complex scene (with distractors)

In the Complex scene, the target was displayed among 2 or 4 distractors that were either Non-aligned with the target, Horizontally-aligned with the target, or Vertically-aligned with the target (Figure 2). Each of the 96 trials (divided into 2 blocks) began with a 500-ms white noise

audio signal indicating that the distractors were going to appear. After 2 seconds during which the distractors were displayed alone, a 400-hz beep was played for 500-ms and immediately after it, the virtual target was displayed among the distractors at one of the 5 possible locations. The distractors and the target disappeared at the same time when the participant logged its response using the pointing tool.

2.4 Data analysis

The R studio software was used for all statistical analyses with version 3.6.1 of R (Team, 2020). In total, 144 response positions (3 blocks x 48 trials) per participant were recorded. Localization performance was assessed separately for the azimuth and the elevation dimensions. For each dimension, localization performance was assessed with regression-based and error-based metrics, analyzed with Linear Mixed Models (LMMs) with the *lmerTest* R-Package (Kuznetsova et al., 2017). The effects were estimated using ANOVAs, and the R-package *emmeans* (Lenth, 2022) was used for post-hoc analyses (version 1.7.4) with Tukey correction.

2.4.1 Regression-based metric: gain and bias

The gain and bias for both dimensions (azimuth and elevation) were estimated with the predictions of the LMMs (response position as a function of target position). The gain was estimated with the predicted slope, while the bias was estimated with the intercepts. An optimal performance would be observed with a gain value of 1.0 and a bias of 0.0°. In contrast, a gain value of 0.0 would reflect a random pattern of responses. In the azimuth dimension, a negative bias would suggest a leftward bias, while in the elevation dimension, a negative bias would suggest an underestimation bias.

For the azimuth and elevation dimensions separately, a first LMM on all response positions was fitted with the Scene (Minimalist scene or Complex scene) and Target position (-20°, 0°, and +20° for the model on azimuth responses, and -25°, 0°, and +25° for the model on elevation responses) as fixed factors to investigate the effect of the presence of the distractors on the pattern of response. Participants were considered as a random factor. A second LMM was fitted only on the response positions in the Complex scene (i.e., with distractors) to investigate the effects of the number of distractors and their spatial disposition relative to the target, with Number (2 or 4 distractors), Disposition (Non-aligned, Horizontally-aligned, or Vertically-aligned) and Target position (same modalities as the first model) as fixed factors, and the participants as a random factor.

2.4.2 Error-based metrics: unsigned error

Elevation (respectively azimuth) unsigned errors were computed as the absolute value of the difference between the target and the response elevation (respectively azimuth). The effect of the presence of distractors (Scene: Minimalist or Complex) on the unsigned errors and the effects of the number of distractors (Number) and their spatial disposition relative to the target (Disposition) were investigated separately with two LMMs. The target position was not included as a fixed factor. The estimated marginal means of the unsigned errors provided by the LMMs were used for post-hoc pairwise comparisons.

3. Results

3.1 Head tracker checks

Since participants were instructed to keep their heads as still as possible, the position of the head tracker was recorded every 200 ms during both the familiarization and the localization tests to check that they respected the instructions. During the familiarization, the maximum distance of the head from its mean position during the whole familiarization was 3.86 ± 1.34 cm (M ± SD), while during the localization tests, the maximum distance of the head from its mean position for each trial was on average 1.4 ± 1.2 cm (M ± SD). In both localization tasks and familiarization, participants thus mainly followed the instructions of keeping their heads still.

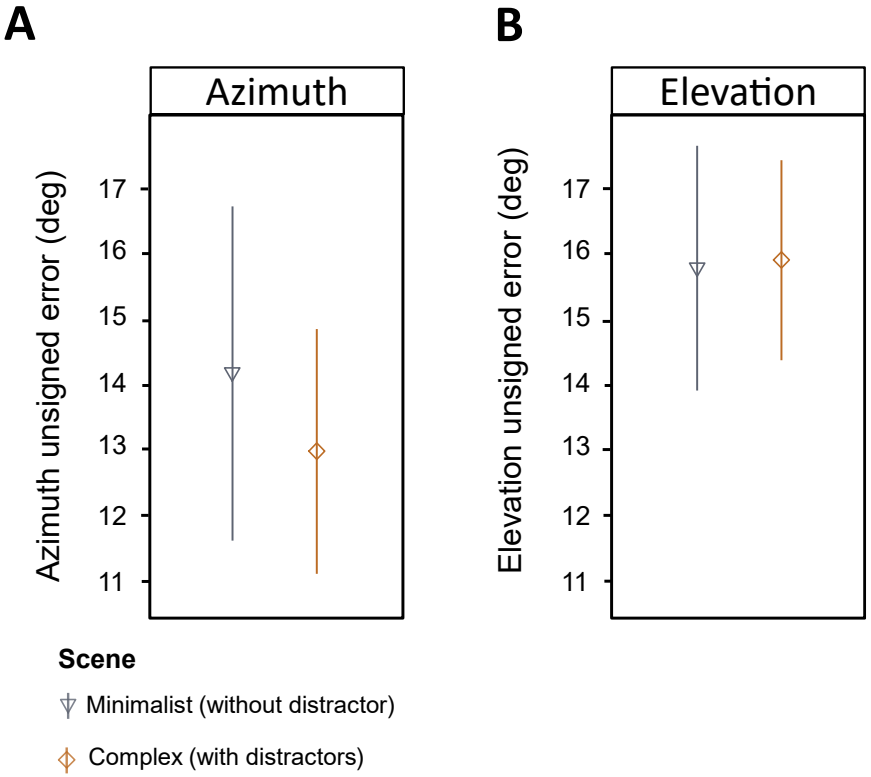
3.2 Effect of the presence of distractors on localization performance

The effect of the presence of distractors on localization abilities, without considering the number of distractors or their spatial disposition, was assessed by comparing the azimuth and elevation localization performance without distractors (Minimalist scene) and with distractors (Complex scene). Regression-based metrics (gain and bias) and error-based metrics (unsigned errors) were analyzed.

3.2.1 Azimuth localization performance with and without distractors

For the error-based metrics, the estimated marginal means of the azimuth unsigned error in the Minimalist and in the Complex scenes are depicted in Figure 4, panel A. The ANOVA did not show a significant effect of the Scene on the azimuth unsigned error [$F(1, 18) = 1.51$, $p = .23$, $\eta_p^2 = 0.08$], with no significant difference between the Minimalist scene (14.2°, 95% CI = [11.3, 17.1]) and the Complex scene (13.0°, 95% CI = [10.9, 15.1]). This result suggests that the accuracy to localize the azimuth was not modulated by the presence of the distractors.

Figure 4. Estimated marginal mean of the unsigned error in the azimuth **(A)** and elevation **(B)** dimensions in the Minimalist scene (gray triangle) and Complex scene (brown diamond), all target positions combined. Error bars show the 95% confidence interval of the estimated marginal means.



For the regression-based metric, the azimuth response positions as a function of the target azimuth in the Minimalist scene and in the Complex scene are depicted in Figure 5, panel A. The ANOVA revealed a significant interaction effect Target azimuth × Scene on the azimuth response position [$F(1, 34.5) = 4.85$, $p = .034$, $\eta_p^2 = 0.12$], suggesting an effect of the Scene (Minimalist or Complex) on the pattern of response in the azimuth dimension. Post-hoc analyses were conducted to specify the effect of the presence of distractors on the azimuth gain and bias. The azimuth gain was estimated with the slope of the model, and the azimuth bias was estimated with the intercept of the model.

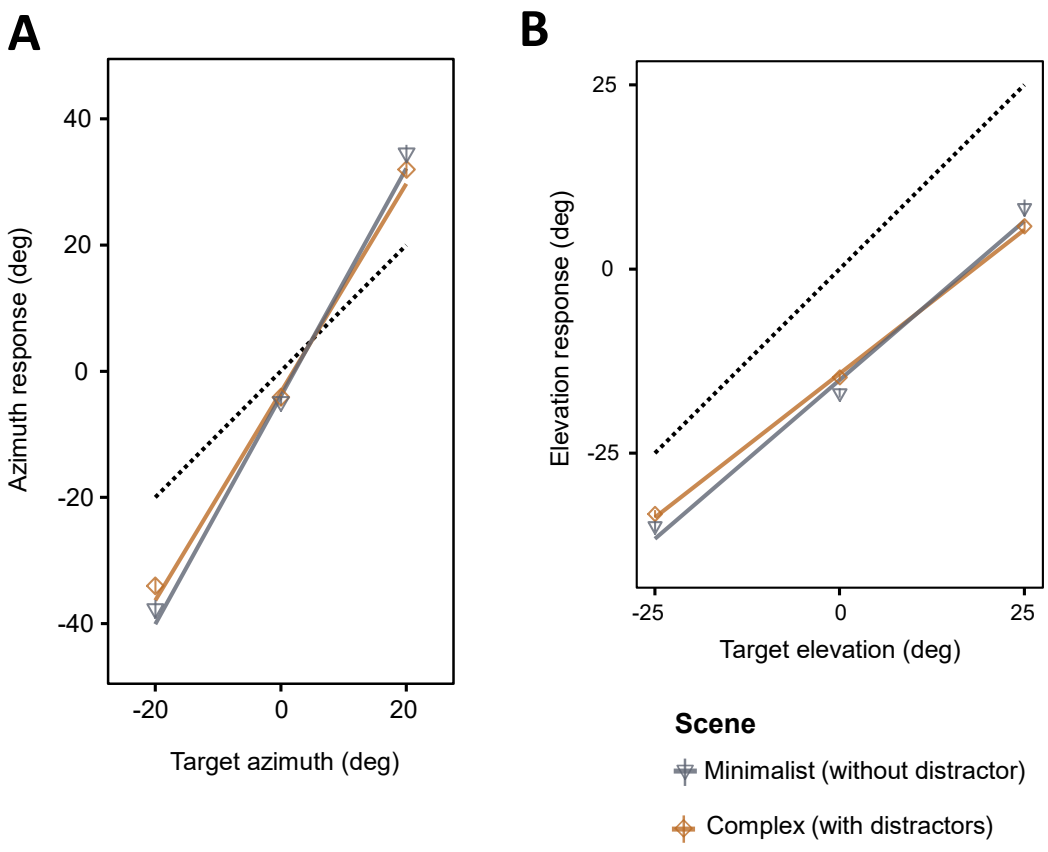
For the gain, in both conditions, the azimuth gain was significantly higher than the optimal gain 1.0 [all $t(18) > 6.62$, all $p < .0001$], which shows a tendency to overestimate the lateral position of the lateral targets. However, the analysis showed that the azimuth gain in the Complex scene (1.65, 95% CI = [1.46, 1.84]) was significantly lower than the azimuth gain in the Minimalist scene (1.81, 95% CI = [1.55, 2.06]), [$t(18) = 2.2$, $p = .0409$] which means that the lateral overestimation pattern was lower when the target was displayed among distractors.

For the bias, the analysis did not show a significant difference between the Minimalist scene (-3.31°, 95% CI = [-5.54, 1.08]) and the Complex scene (-3.97°, 95% CI = [-7.15, -0.79]). However,

the azimuth bias was significantly lower than the optimal bias of 0° in both conditions [all $t(18) > 2.45$, all $p < .0248$)], suggesting a leftward tendency whatever the presence of distractors.

To sum up, these results show that participants were able to perceive the azimuth location of the target regardless of the presence of distractors. They also show a lateral overestimation pattern and a slight tendency to judge the azimuth at a more left location.

Figure 5. Mean response position as a function of the target position for the azimuth **(A)** and elevation **(B)** dimensions. Black dashed lines indicate optimal performance with gain = 1.0 and bias = 0°. **(A)** Azimuth response position as a function of the target azimuth in the Minimalist scene (gray triangle) and Complex scene (brown diamond). Error bars show the standard error of the azimuth response position Solid lines represent the azimuth gains (estimated by the slopes provided by the LMM) in the Minimalist scene (gray) and Complex scene (brown). **(B)** Elevation response position as a function of the target elevation in the Minimalist scene (gray triangle) and Complex scene (brown diamond). Error bars show the standard error of the elevation response position. Solid lines represent the elevation gains (estimated by the slopes provided by the LMM) in the Minimalist scene (gray) and Complex scene (brown).



3.2.2 Elevation localization performance with and without distractors

For the error-based metrics, the estimated marginal means of the elevation unsigned error in the Minimalist scene and in the Complex scene are depicted in Figure 4, panel B. The ANOVA did not show a significant effect of the Scene on the elevation unsigned error [$F(1, 18) = .03$, $p = .86$, $\eta_p^2 = 0.0017$], with no difference between the Minimalist scene (15.8°, (95%CI = [13.6, 18.0])

and the Complex scene (15.9°, 95%CI = [14.1, 17.7]). This result suggests that the accuracy to localize the elevation did not seem to be modulated by the presence of the distractors.

For the regression-based metric, the elevation response positions in the Minimalist scene (without distractor) and in the Complex scene (with distractors) are depicted in Figure 5, panel B. The interaction effect Target elevation × Scene was not significant [$F(1, 17.998) = 1.85$, $p = .19$, $\eta_p^2 = 0.09$], suggesting that the pattern of response in the elevation dimension was comparable when the target was displayed alone or among distractors. For a descriptive purpose, the elevation gains and bias were estimated with the predictions of the LMM in the Minimalist and Complex scenes. The elevation gain in the Minimalist scene (0.86, 95% CI = [0.68, 1.04], $t(18) = 1.62$, $p = .12$) was not significantly different from the optimal elevation gain 1.0, while in the Complex scene (0.78, 95%CI = [0.65, 0.91], $t(18) = 3.63$, $p = .0019$) the elevation gain was significantly lower than this optimal value. This result suggests an elevation compression pattern when the target had to be localized among distractors, although the pattern of response in the elevation dimension does not seem to be drastically different, since the elevation gain when the target was displayed among distractors was not significantly lower than when the target was displayed alone.

For the elevation bias, it was significantly lower than the optimal bias of 0° both in the Minimalist scene (-15.1°, (95%CI = [-17.5, -12.6]) and in the Complex scene (-14.2°, (95% CI = [-16.5, -11.9]) [all $t(18) > 12.7$, all $p < .0001$], which shows an underestimation of the elevation of the target.

Overall, without considering the number of distractors and their spatial disposition, participants successfully localized the target with the SSD in the azimuth and elevation dimensions, even when it was displayed in a complex scene containing distractors. The azimuth localization performance was characterized by an overestimation pattern and a slight leftward bias, while the elevation was underestimated, with a slight elevation compression pattern observed when distractors were displayed.

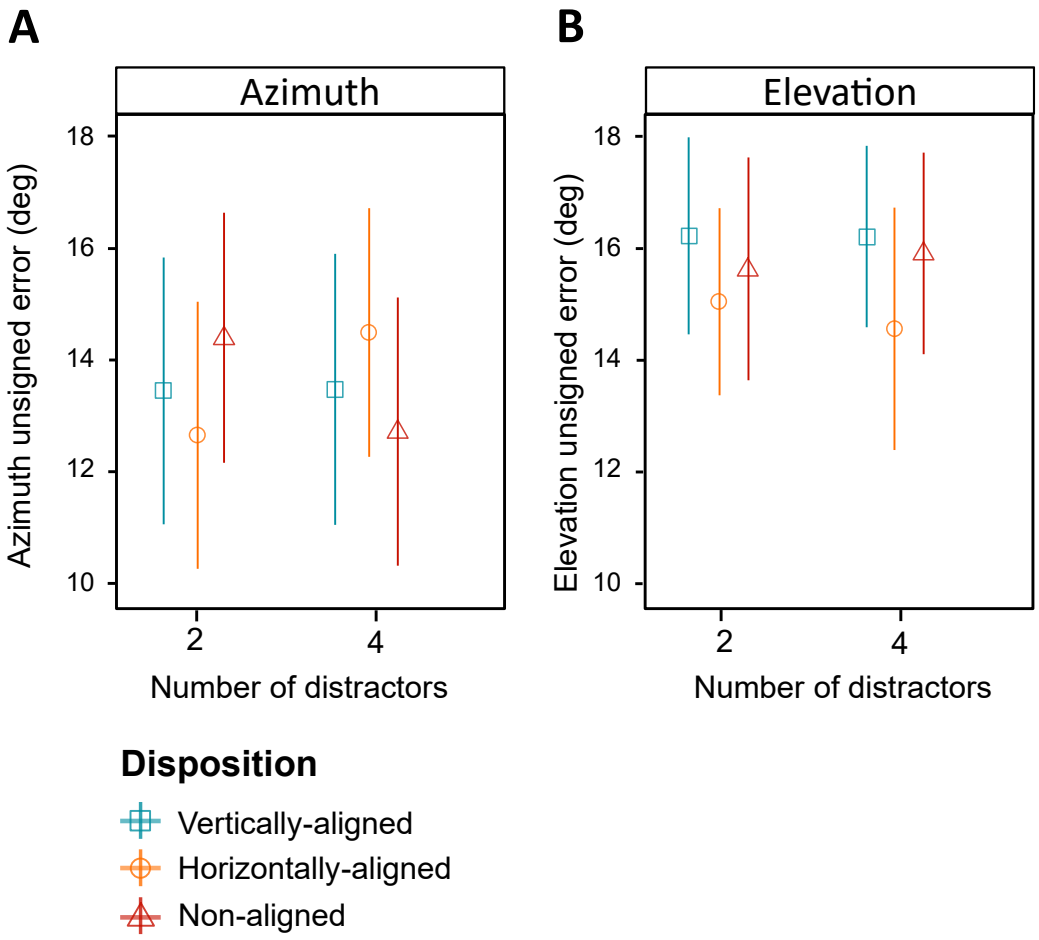
3.3 Effect of the number of distractors and their spatial disposition on localization performance

In order to investigate the effect of the number of distractors (Number: 2 distractors, or 4 distractors) and their spatial disposition relative to the target location (Disposition: Non-aligned, Horizontally-aligned, or Vertically-aligned) another statistical model was used that only takes into account the trials for which the target was displayed among distractors (Complex scene only). Regression-based metrics (gain and bias computed based on the response positions) and error-based metrics (unsigned errors) were analyzed.

3.3.1 Azimuth localization performance in the complex scene

For the error-based metrics, the estimated marginal means of the azimuth unsigned error in the 6 experimental conditions are depicted in Figure 6, panel A. The ANOVA showed that the interaction effect Number × Disposition [$F(2, 33.814) = 4.0$, $p = .0275$, $\eta_p^2 = 0.19$] was significant, however post-hoc analyses with Tukey correction did not show any significant differences [all $t(18) < 1.996$, all $p > .102$] between the 6 experimental conditions. Therefore, the accuracy to localize the azimuth did not seem to be modulated by the number of distractors or by their spatial disposition relative to the target.

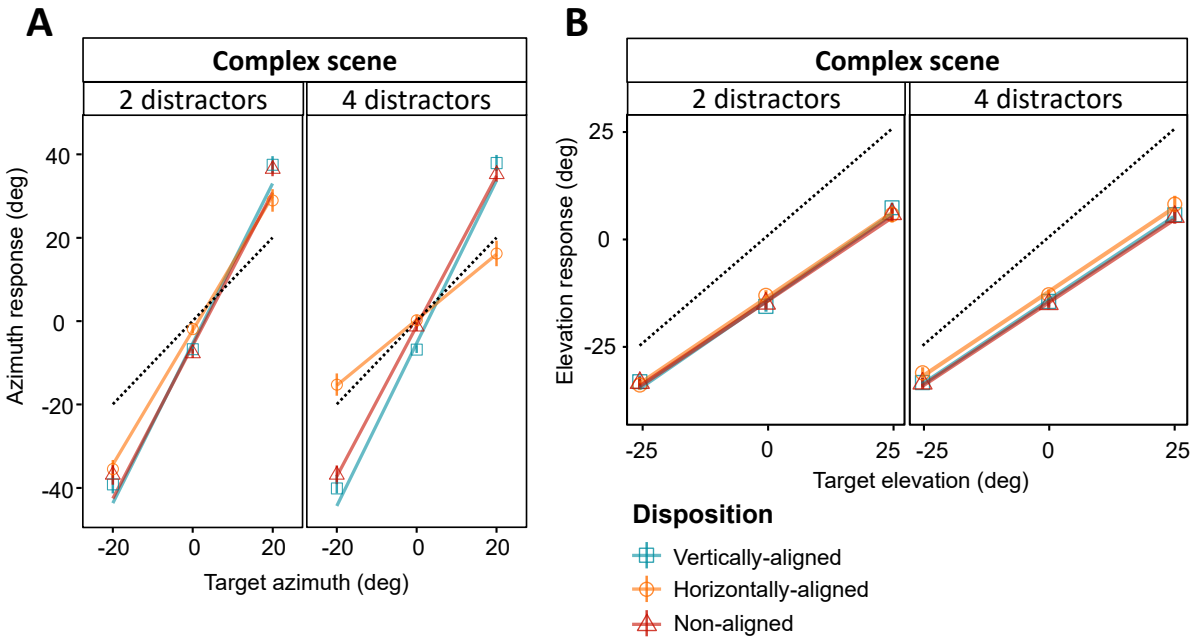
Figure 6. Estimated marginal mean of the unsigned error in the azimuth **(A)** and elevation **(B)** dimensions in the Complex scene. Error bars show the 95% confidence interval of the estimated marginal means. **(A)** Estimated marginal means of the azimuth unsigned error as a function of the Number of distractors (2 or 4) when the distractors are Vertically-aligned (blue square), Horizontally-aligned (orange circle), and Non-aligned with the target (red triangle) **(B)** Estimated marginal means of the elevation unsigned error as a function of the Number of distractors (2 or 4) when the distractors are Vertically-aligned (blue square), Horizontally-aligned (orange circle), and Non-aligned with the target (red triangle).



For the regression-based metric, the azimuth response positions in the Complex scene are depicted in Figure 7, panel A. The azimuth gains and bias were estimated with the predictions of

the LMM (respectively the slopes and intercepts) and are summarized in Table 1 in the 6 experimental conditions (with 95% confidence interval).

Figure 7. Mean response position as a function of the target position for the azimuth **(A)** and elevation **(B)** dimensions in the Complex scene. Black dashed lines indicate optimal performance with gain = 1.0 and bias = 0°. **(A)** Azimuth response position as a function of the target azimuth, and the Number of distractors (2 distractors in the left figure; 4 distractors in the right figure). Symbols represent the mean azimuth response positions when the distractors are Vertically-aligned (blue square), Horizontally-aligned (orange circle), and Non-aligned with the target (red triangle). Solid lines represent the azimuth gains (estimated by the slopes provided by the LMM) when the distractors are Vertically-aligned (blue), Horizontally-aligned (orange), and Non-aligned with the target (red). **(B)** Elevation response position as a function of the target elevation, and the Number of distractors (2 distractors: left figure; 4 distractors: right figure). Symbols represent the mean elevation response positions when the distractors are Vertically-aligned (blue square), Horizontally-aligned (orange circle), and Non-aligned with the target (red triangle). Solid lines represent the elevation gains (estimated by the slopes provided by the LMM) when the distractors are Vertically-aligned (blue), Horizontally-aligned (orange), and Non-aligned with the target (red).



The azimuth gains in the 6 experimental conditions were significantly different from the optimal gain 1.0 [all $t(18) > 6.37$, all $p < .0001$], however two distinct patterns of response were measured depending on the number of distractors and their spatial disposition. Except in the experimental condition where 4 horizontally aligned distractors were displayed (Horizontally-aligned condition), the azimuth gains in the five other experimental conditions showed a tendency to overestimate the lateral position of lateral targets (azimuth gains between 1.61 and 1.95). In contrast, when 4 distractors were horizontally aligned with the target (Horizontally-aligned condition), a reverse pattern of azimuth compression was observed. As shown in Figure 7, panel A (right figure), the azimuth gain drastically decreased (azimuth gain of 0.79, lower than the optimal gain 1.0) which shows a lateral underestimation pattern. This decrease did not seem to reflect a

random pattern of response, since the azimuth gain in this condition was still significantly higher than 0.0 [$t(18) = 6.424$, $p < .0001$]. Therefore, participants tended to underestimate the lateral position only when 4 distractors were horizontally aligned with the target.

Table 1. Azimuth gain and bias for each Disposition condition (Non-aligned, Vertically-aligned, Horizontally-aligned) and Number condition (2 distractors, 4 distractors). 95% confidence interval is given in brackets. The (*) symbol indicates a significant difference between the azimuth gain and the optimal gain 1.0, or between the azimuth bias and the optimal bias 0.0°.

		Non-aligned	**Vertically-aligned**	**Horizontally-aligned**
2 distractors	Azimuth gain	1.84 [1.59, 2.08] (*)	1.92 [1.66, 2.17] (*)	1.61 [1.38, 1.84] (*)
	Azimuth bias	-5.90° [-8.68, -3.12] (*)	-5.37° [-8.59, -2.14] (*)	-2.26° [-4.76, 0.23]
4 distractors	Azimuth gain	1.80 [1.55, 2.06] (*)	1.95 [1.71, 2.19] (*)	0.79 [0.53, 1.05] (*)
	Azimuth bias	-1.24° [-4.50, 2.01]	-5.36° [-8.70, -2.02] (*)	0.26° [-2.93, 3.45]

This heterogeneity in the results is confirmed by an ANOVA that raised a significant interaction effect Target azimuth × Number × Disposition [$F(2, 69.004) = 13.88$, $p < .0001$, $\eta_p^2 = 0.29$]. Post-hoc analyses were thus conducted to specify the interaction effect between Disposition (Non-aligned, Horizontally-aligned, or Vertically-aligned) and Number (2 distractors or 4 distractors) on the azimuth gains and bias.

For this purpose, we first investigated if the increase of the number of distractors resulted in a modulation of the azimuth gain depending on the spatial disposition of the distractors relative to the target (i.e., effect of the Number of distractors on the azimuth gain for the three Disposition conditions, separately). The analysis confirmed that the decrease in the azimuth gain observed when the number of distractors increased from 2 to 4 was only present when the distractors were horizontally aligned with the target (from 1.61 with 2 distractors to 0.79 with 4 distractors), [$t(18) = 6.29$, $p < .0001$].

Secondly, we investigated if the spatial disposition of the distractors had an influence on the azimuth gain with 2 or 4 distractors (i.e., effect of the Disposition of the distractors on the azimuth gain depending on the Number of distractors). This analysis confirmed the specificity of the configuration with 4 horizontally aligned distractors. With 4 distractors, the azimuth gain in the Horizontally-aligned disposition (0.79) was significantly lower than with the two other spatial dispositions (Non-aligned: 1.80, Vertically-aligned: 1.95), [all $t(18) > 7.58$, $p < .0001$]. In contrast, with 2 distractors, the azimuth gain in the Horizontally-aligned condition was only marginally lower than when the distractors were vertically aligned with the target (1.92, [$t(18) = 2.529$, $p = .0523$]), while no difference was found with the azimuth gain in the Non-aligned condition (1.84, [$t(18) = 1.89$, $p = .171$]).

For the azimuth bias (Table 1), post-hoc analyses show that a leftward bias was only observed when the distractors were vertically aligned with the target (Vertically-aligned with 2 distractors: -5.37°, Vertically-aligned with 4 distractors: -5.36°), and when 2 distractors were non-aligned with the target (Non-aligned with 2 distractors: -5.9°) [all $t(18) > 3.14$, all $p < .0056$]. In the remaining experimental conditions, there was no significant shift to the right or to the left.

To sum up the results on the azimuth, the accuracy measured with the unsigned error suggests that when the target was displayed among distractors, participants localized the azimuth of the target with a similar accuracy, whatever the spatial disposition of the distractors and whatever their number. However, the azimuth localization pattern measured with the azimuth gains reveals a lateral underestimation of the azimuth when 4 distractors were horizontally aligned with the target. In contrast, in all the other configurations, an overestimation pattern similar to the one observed without distractors was measured. A slight tendency to judge the azimuth at a more left location was systematically observed when the distractors were vertically aligned with the target, and also when 2 distractors were displayed but not aligned with the target.

3.3.2 Elevation localization performance in the complex scene

For the error-based metrics in the elevation dimension, the estimated marginal means of the elevation unsigned error in the 6 experimental conditions are depicted in Figure 6, panel B. No significant effect of Number (2 distractors or 4 distractors) or Disposition (Non-aligned, Horizontally-aligned, or Vertically-aligned) nor interaction effect Number × Disposition was observed in the analysis. It suggests that the number of distractors and their spatial disposition relative to the target did not modulate the accuracy to localize the elevation of the target.

For the regression-based metric, the elevation response positions in the Complex scene are depicted in Figure 7, panel B, as a function of the number of distractors (Number) and their spatial

disposition relative to the target (Disposition). The corresponding values of the elevation gains and the bias (with the 95% confidence interval) are summarized in Table 2 for the 6 experimental conditions. The gain and bias of the elevation were compared for the 6 experimental conditions to, respectively, the optimal gain 1.0 and the optimal bias 0.0°. For the elevation gains, a compression pattern was measured with the 6 experimental conditions since all the elevation gains were significantly lower than the optimal gain 1.0 [all $t(18) > 2.38$, all $p < .0286$]. For the elevation bias, the underestimation was also observed in the 6 experimental conditions with elevation bias significantly lower than 0° [all $t(18) > 7.868$, all $p < .0001$].

Table 2. Elevation gain and bias for each Disposition condition (Non-aligned, Vertically-aligned, Horizontally-aligned) and Number condition (2 distractors, 4 distractors). 95% confidence interval is given in brackets. The (*) symbol indicates a significant difference between the elevation gain and the optimal gain 1.0, or between the elevation bias and the optimal bias 0.0°.

		Non-aligned	**Vertically-aligned**	**Horizontally-aligned**
2 distractors	Elevation gain	0.78 [0.62, 0.94] (*)	0.80 [0.67, 0.94] (*)	0.79 [0.63, 0.94] (*)
	Elevation bias	-14.7° [-17.1, -12.32] (*)	-14.6° [-17.3, -11.85] (*)	-13.9° [-16.5, 11.29] (*)
4 distractors	Elevation gain	0.77 [0.63, 0.90] (*)	0.77 [0.64, 0.91] (*)	0.78 [0.58, 0.97] (*)
	Elevation bias	-15.0° [-18.2, -11.73] (*)	-14.5° [-16.9, -12.03] (*)	-12.5° [-15.9, -9.18] (*)

The number of distractors and their spatial disposition relative to the target did not influence the pattern of response in the elevation dimension since the ANOVA on the elevation response position did not raise any significant effect except the main effect of the Target elevation [$F(1, 17.96) = 166.94$, $p < .0001$, $\eta_p^2 = 0.9$].

To sum up, concerning the elevation, when the target was displayed among distractors, participants localized with a similar accuracy the elevation of the target, whatever the spatial disposition of the distractors and whatever their number. In all the conditions, the perception of the elevation of the target was similarly compressed and underestimated.

4. Discussion

In this work, we investigated the ability to use a visual-to-auditory SSD to localize an object displayed among other irrelevant objects considered as distractors. After a brief familiarization with the SSD principles, the early-stage abilities were assessed with a blindfolded pointing localization task in a virtual environment. The effect of the presence of distractors on localization abilities was assessed by comparing localization performance with and without distractors, while the effect of the spatial arrangement of the scene was investigated through the manipulation of the number of distractors and their spatial disposition relative to the target.

4.1 Localization abilities in a minimalist scene

4.1.1 Sound spatialization is an effective encoding scheme for the azimuth dimension although an overestimation pattern is observed

In the minimalist scene without distractors, localization performance for the azimuth dimension showed a lateral overestimation pattern (azimuth gain of 1.81) and a slight leftward bias (-3.31°). In the SSD domain, the task described in the current study is comparable to the one presented in Bordeau et al. (2023) which uses the same SSD encoding scheme (called the Monotonic encoding in their study) and a similar experimental set-up (pointing task and familiarization method). In this previous study, they did not observe a leftward bias (-1.8° but not significant), however a leftward shift was measured with other tested encoding schemes. The reasons for those slight leftward bias are unclear, but as explained in Bordeau et al. (2023), they could come from a perceptive or proprioceptive bias. However, in line with the current study, a lateral overestimation pattern was observed in the Monotonic encoding condition with an azimuth gain of 1.23. Although present, this previous lateral overestimation pattern seems thus lower than the one reported in the current study (1.81). It could be the result of a smaller tested azimuth range in the current study ([-20°, +20°]) compared to the previous one ([-40°, +40°]).

This overestimation pattern of lateral sound sources has also been reported in auditory localization studies with real sound sources (Oldfield & Parker, 1984), simulated sound sources using non-individualized HRTFs (Wenzel et al., 1993) or in virtual environments (Ahrens et al., 2019). Oldfield & Parker (1984) propose that this overestimation pattern could be explained by the cone of confusion, which is a set of sound source locations for which the spatial binaural cues are very similar and thus difficult to distinguish from each other to determine the true location. These positions are symmetrically placed around the transversal axis. This could therefore result in localization confusion, which is known to be increased with non-individualized HRTFs such as

those used in the current study. It is thus well possible that this perceptive bias occurred, even if participants were told that the target could only be located in the front space.

In our study, the azimuth unsigned error without distractors of 14.2° is comparable to the range of 15° to 19° measured after the familiarization in Bordeau et al. (2023). Although azimuth error has also been measured with other SSDs, the values are difficult to compare since the tasks were different. For instance, using pointing tasks on a table, Hanneton et al. (2010) measured an angular error of about 5° with the Vibe SSD, and Commère & Rouat (2023) measured azimuth errors between about 8° and 45° with their SSD that conveys azimuth location with stereo panning. Using a body pointing task (i.e., facing the target) with an SSD, Scalvini et al. (2022) measured an azimuth error of about 7°, which is two times smaller than in the current study but which could be easily explained by the different pointing method. Mendonça et al. (2013) measured a similar azimuth unsigned error of about 15° with simulated sound sources spatialized using the same non-individualized HRTFs database as in this study.

Overall, a strong laterality overestimation pattern was observed when the participants had to localize the target in a minimalist scene without distractors. This overestimation pattern is consistent with previous auditory localization experiments conducted with real and simulated sound sources with non-individualized HRTFs. When using an SSD, the lateral overestimation of an obstacle could result in a collision. However, in a real context of use, SSD users would have the possibility to move their heads in order to improve their perception by placing the target in the median axis, such as with experiments using a body or a head-pointing method.

4.1.2 Pitch modulation is an effective acoustic cue for the elevation dimension

Elevation localization performance measured with gain and bias depicts a good ability to discriminate the three tested elevations. The elevation gain of 0.86 is comparable to the optimal gain 1.0, although an underestimation bias of -15.1° is measured. In Bordeau et al. (2023), the elevation gain with the similar SSD encoding scheme (i.e., Monotonic encoding) was 1.015, which was also comparable to the optimal gain 1.0. The underestimation bias measured in the current study was also observed in Bordeau et al. (2023) and seems comparable (-14.15°). As explained in Bordeau et al. (2023), it is possible that this underestimation bias comes from the high position of the head tracker in the front head of the participants that was associated with the virtual camera. It resulted in a field of view where the 0° elevation coordinate corresponds to the axis straight ahead of the head tracker, which is higher than the ears' level.

The unsigned elevation error observed in the current study was about 16°, which is comparable to the range of 17° to 24° after the familiarization in Bordeau et al. (2023). Using the

Synaestheatre SSD (Hamilton-Fletcher et al., 2016) that conveys elevation and azimuth only using spatial cues based on non-individualized HRTFs, Richardson et al. (2019) measured an elevation discrimination score of 14°, which is also similar to what we found. As a comparison in the framework of auditory localization, Mendonça et al. (2013) measured an elevation error of about 25° after the participants were familiarized with sounds spatialized using the same HRTFs database as used in this study. However, in Mendonça et al. (2013), they tested higher elevation locations (elevation locations ranging from 0° to +90°), which probably resulted in higher localization error since elevation localization abilities are lower for high elevations (Makous & Middlebrooks, 1990).

Overall, in a minimalist scene without distractors, the brief audio-motor familiarization with the SSD encoding scheme was sufficient to enable participants to localize the elevation of the target based on spectral information resulting from the pitch modulation in the elevation encoding scheme. Although the elevation was strongly underestimated, it probably arises from the spatial disparity between the egocentric spatial mental representation and the head tracker location. Therefore, it could be reduced with SSD practice through a recalibration of the auditory spatial perception, as in the ventriloquism effect (Howard, 1966).

4.2 Localization abilities in a complex scene

To assess localization abilities with the SSD in a complex scene, we tested the localization performance with a target displayed among distractors. The distractors were either aligned with the target (horizontally or vertically) or non-aligned. Overall, localization performance in the elevation dimension was preserved with the distractors, although the presence of distractors modulated the pattern of response in the azimuth dimension.

4.2.1 Causes of the azimuth underestimation pattern when 4 distractors are horizontally aligned

The laterality overestimation pattern observed without distractors was also present when the target was displayed among distractors, with an azimuth gain of 1.65, although it was lower compared to the azimuth gain of 1.81 without distractors. However, this decrease in the azimuth gain observed with distractors mainly comes from a specific configuration where 4 distractors were horizontally aligned with the target. In this condition, a reverse pattern of lateral underestimation was observed, with an azimuth gain of 0.79.

This drastic decrease of the azimuth gain resulting in a lateral underestimation pattern may reflect a localization impairment for the azimuth dimension since it shows a bias toward the median axis (i.e., 0° azimuth) which can be considered as a pattern where only random responses are given. However, although it is getting closer to 0, it is still significantly and largely higher. Therefore, the

lateral underestimation pattern in this spatial disposition may reflect difficulties in localizing the azimuth of the target, but not a complete inability. The potential reasons for these difficulties are discussed below.

4.2.1.1 Is it the result of the signal-to-noise ratio decrease?

A first explanation could be the decrease in the signal-to-noise ratio (SNR) as the number of distractors increases from 2 to 4. If we assimilate the target as the signal and the distractors as the noise, increasing the number of distractors resulted in a decrease in the SNR. In the framework of auditory localization, it is well known that azimuth localization abilities are impaired when the SNR decreases (Kerber & Seeber, 2012; Lorenzi et al., 1999). For instance, a decrease in the SNR has been found to be associated both with a decrease in azimuth gains resulting in a lateral underestimation pattern in Kerber & Seeber (2012), and with an increase in the azimuth angular error in Lorenzi et al. (1999).

However, we did not observe a systematic decrease in the azimuth gain when distractors were displayed. Indeed, when the distractors were non-aligned or vertically aligned with the target, localization performance was not impaired by the increase in the number of distractors. Therefore, the decrease in the SNR does not seem to be the only explanation for this localization impairment. It suggests that it is not only the number of distractors that results in localization impairment for the azimuth but also their spatial disposition relative to the target.

4.2.1.2 Is it the angular separation between the target and the distractors that is too low?

A second explanation could be that the target and the distractors were separated by a too small angle, preventing the perceptive segregation of the target from the distractors. It is known that in the case of two real sound sources played simultaneously, the angular separation between them modulates localization abilities (Best et al., 2004; Perrott, 1984), also known as the minimal audible angle (MAA). Best et al. (2004) measured a minimum audible angle of about 20° when two spatialized broadband sounds were located laterally (between 22.5° and 45°), while Perrott (1984) measured a MAA between 10° and 16° when two low tones were played laterally (25° or 40° in azimuth). Therefore, when the target was horizontally aligned with 4 distractors in the current study, the angular separation of 20° in azimuth between the target and the closest distractors could have been too low to be easily perceived. However, this angular separation was the same as the condition when 2 horizontally aligned distractors were displayed, and no localization impairments were found in this spatial disposition. Moreover, when the distractors were non-aligned with the target, the closest distractor was separated by 10° in azimuth from the target, which is even lower, but still no localization impairments were measured in this spatial disposition.

Unlike in MAA experiments, more than two "sound sources" were played in the current study, since at least two distractors were displayed simultaneously in addition to the target. The number of simultaneous sound sources is known to influence the spatial separation required to localize them (Eramudugolla et al., 2005; Kawashima & Sato, 2015; Zhong & Yost, 2017 ; for review, see Kwak & Han, 2020). For instance, the maximum number of sound sources that could be separately perceived was about 3 for real tones ranging from 313 to 5051 Hz (Zhong & Yost, 2017), and between 4 and 5 for spatialized environmental sounds using non-individualized HRTFs (Kawashima & Sato, 2015, Eramudugolla et al., 2005), with sound sources separated by at least 30° in azimuth in the three mentioned studies. Since, in the current study, the target and each of the four distractors were separated by at least 20° in the azimuth dimension when they were horizontally aligned, it is unlikely that all five objects were perceived as distinct objects through the soundscape. However, if the fusion of the distractors was the reason for the observed localization impairments, it should have been also observed when the distractors and the target were vertically aligned or not aligned. In contrast, azimuth localization performance was not impaired in these spatial dispositions, suggesting that there is another explanation behind it.

4.2.1.3 Is it the lateral eccentricity of the target?

The unique spatial configuration with 4 horizontally aligned distractors could be mentioned as a third possible reason for this special observed effect since the distractors were spatially distributed around the target on both hemifield (left and right) only with this configuration, whatever the location of the target (median axis or lateral location). On the contrary, with the same spatial disposition with only two distractors, the lateral target was displayed at a more eccentric lateral location than the distractors. As discussed in the previous section, in the case where several sounds are played from different locations, the spatial separation of the sound sources prevents the masking effect, which is known as the spatial release from masking (Kawashima & Sato, 2015). Therefore, with four distractors, it is possible that the spatial release from masking was not possible due to the eccentric distractors located at ± 40°. On the other hand, when the distractors were neither horizontally nor vertically aligned with the target, they were also spatially distributed around the target, and we did not observe azimuth localization impairments in this spatial disposition, even with 4 distractors.

In the cocktail party problem in auditory experiments, the spatial release from masking relying on binaural cues seems robust even when the maskers are spatially distributed around the acoustic signal in both hemifields as shown in Hawley et al. (2004). Therefore, at first glance, we should have observed a comparable masking effect in both the horizontally aligned and the non-aligned dispositions in the current study. However, the previously mentioned study used broadband

sounds (speech or envelope-modulated noise), while in the current study, the SSD soundscapes coming from the sonification of each object (target or distractor) were narrowband. As a consequence, when distractors were located at distinct elevations from the target, the target was associated with a distinctive spectral signature that could be used for localization. Therefore, if such masking occurred when the distractors were horizontally aligned, it was not only the result of a decrease in the SNR, nor only the result of binaural cues that were not reliable enough for a spatial release from masking occurs.

4.2.1.4 Is it a masking effect due to the bandwidth of the auditory filters?

The fourth possible explanation is based on a pitch masking effect. Since the elevation encoding scheme of the SSD includes the modulation of the pitch, the configuration when the distractors and the target are horizontally aligned results in a similar narrowband frequency spectrum associated with each object (Figure 3). In contrast, the reported capacity to separate between 4 and 5 spatialized environmental sounds using non-individualized HRTFs found in the previous mentioned studies of Kawashima & Sato (2015) and Eramudugolla et al. (2005) was observed using wider-band sounds, and in Zhong & Yost (2017) the capacity to separate 3 sound sources was found with tones based on frequencies at least 106 Hz apart and not constituting harmonic series. In the current study, it is thus highly possible that the spectral similarity between distractors and the target has caused higher difficulties in segregating the different objects through the soundscape. The frequency composition of a soundscape is known to influence the sound source segregation abilities due to the auditory filters emerging from the cochlear tonotopy (Bregman, 1990), also known as critical bands (Glasberg & Moore, 1990; Zwicker, 1961). This phenomenon implies that the segregation of two narrowband sounds is impaired when they share frequency components on a given bandwidth called the critical band. In other words, if the frequency spectrum of the sound masker is too close to the frequency of the tone, a masking effect occurs, which roughly consists of the fusion of the sound masker with the tone.

As in the current study, Ambard et al. (2015) observed a bias toward the median axis when two horizontally aligned objects were simultaneously perceived through the SSD soundscape. Therefore, a masking effect probably also occurred when the target and the distractors were horizontally aligned, thus sharing the same spectral composition. In this configuration, the perceptive segregation of the target and the distractors could only rely on the spatial binaural cues provided by the spatialization with non-individualized HRTFs. However, in the current work, the localization impairments do not seem to be entirely caused by the spectral similarity resulting from the spatial disposition since the localization impairments were observed only when 4 horizontally aligned distractors were displayed, but not 2. Therefore, it seems that the underestimation pattern

was caused by a pitch masking effect that occurs when the SNR within a narrow frequency range reaches a sufficiently low threshold.

Concerning the leftward bias, it was not systematic when the scene was complex but only when the distractors were vertically aligned with the target, or when they were not aligned but with only two distractors. The presence of a systematic leftward bias when the distractors were vertically aligned with the target is consistent with the leftward bias observed in the minimalist scene (without distractors) since the azimuth location of the distractors and the target was the same in this complex scene disposition, although it is unclear whether it has a perceptive or proprioceptive origin.

Taken together, the results concerning the azimuth localization abilities in the complex scene show that spatial binaural cues can be efficiently used even when no target-specific spectral signature is provided, although this capacity seems to quickly decrease when many distractors are located on the same elevation, resulting in a pitch masking effect when the SNR is too low. However, if such a masking occurred, it was not a full masking effect since participants could still perceive that the target was located at a lateral location, but underestimated it. On the other hand, when measuring the accuracy to localize the azimuth of the target using the azimuth unsigned error, there was no effect of the spatial disposition of the complex scene, from which one could conclude that the reverse pattern of lateral underestimation does not result in a lower accuracy than the lateral overestimation usually observed. However, an underestimation pattern could depict a tendency to reply more randomly in comparison to an overestimation pattern, which, on the contrary, can reflect a higher sensitivity. If such a masking effect occurs, it means that it should be difficult to detect the apparition of the target among the distractors without the 440 Hz beep signal preceding the target display. It could be assessed in a future study using a detection task where a target appearing at a random time among distractors had to be detected, then localized, similarly to (Eramudugolla et al., 2005).

4.2.2 The pitch modulation preserves the elevation localization performance in a complex scene

Localization performance for the elevation dimension was not strongly impaired by the presence of distractors, whatever their number or their spatial disposition. However, the presence of distractors resulted in a compressive bias, as suggested by an elevation gain significantly lower than the optimal gain (elevation gain of 0.78), which was not observed when the target was displayed alone, resulting in an elevation gain of 0.86 (but not significantly lower than 1.0). The accuracy in elevation as well as the downward bias were similar in the complex and minimalist scenes (respectively 15.9° and 15.8° for the unsigned error and -14° and -15° for the downward bias).

The SSD used in this study conveys the elevation dimension through spatial cues provided by the HRTFs, but mainly with the modulation of pitch in the frequency range [250, 1492 Hz]. The ability to segregate sound sources if they do not share the same spectral composition that we measured in the current study is in line with what Zhong & Yost (2017) found in auditory scene analysis, where the segregation of speech sounds and tones has been found to be based not only on a spatial processing but also on other acoustic features such as pitch or timbre. In their study, multiple speech sounds or tones could be detected even if they were played from the same loudspeaker (i.e., the same location). However, the correct identification of the number of simultaneously played sounds rarely exceeded three, even if none of the tones constituted a harmonic series and if the used frequencies were separated by at least 106 Hz. In the current study, although it is difficult to assess whether participants were able to perceive five distinct objects in the most complex configuration (with four distractors), they were clearly able to segregate the elevation location of the target.

The range of frequency used in the SSD used for this study ([250 Hz, 1492 Hz]) covers a range of approximately 30 successive Equivalent Rectangular Bands (ERBs) as defined in Moore and Glasberg (1990). Theoretically, this would limit the elevation segregation abilities to a maximum of 30 distinct elevation locations, although in practice, the segregation abilities of sound sources are more in the range of 3 to 5 simultaneous sound sources (Brungart et al., 2005; Zhong & Yost, 2017). However, segregating two (or more) objects located at distinct elevations depends on their proximity to each other (i.e., depends on the elevation's angular separation). In the current study, only 5 elevation locations were simultaneously occupied at maximum in the situation when 4 distractors were vertically aligned with the target. In this spatial disposition, each object was separated by 12.5° in elevation. Although considering the size of the distractors and the target on the video frame with a vertical resolution of 120 pixels, we could estimate that each of the five objects was separated by at least two independent ERBs as defined in Moore and Glasberg (1990) and therefore characterized by a specific spectral signature.

Overall, pitch modulation seems an efficient acoustic cue to be used in SSD encoding schemes since it is quickly interpretable to localize the elevation of an object in a complex scene where it has to be segregated from other irrelevant objects, which is a common situation in the context of pedestrian trips. Due to the auditory filtering of the auditory system, the scene segregation abilities relying on this spectral cue depend on both the frequency range and resolution. The current SSD, using a frequency range of 250–1492 Hz and a frequency resolution of 120 frequencies distributed in about 30 separated ERBs, seems sufficiently reliable to perceptively isolate an object from four others.

4.3 Implications for SSD design for the blind

Although the virtual environment used in the current study is still far from a rich real environment like a crowded street, the proposed protocol allows for the assessment of target localization abilities within scenes of increasing complexity. The abilities to localize the lateral position of the virtual object seemed impaired when irrelevant objects were displayed along the horizontal field of view. As discussed above, this impairment is probably due to the difficulty of segregating all the objects of the visual scene through the soundscape. In a real context of SSD use, this situation would arise when multiple objects are located at distinct azimuth locations but of similar heights and at a similar distance from the camera, such as a line of small posts in the street that has to be crossed.

It raises the question of the size of the field that is transmitted through the soundscape of the SSD. In the current study, the horizontal field of view of the virtual camera was 90°, and the objects (target and distractors) were located between -40° and +40° in azimuth. If the visual scene is too difficult to be accurately segregated by the users, there is no interest in conveying this large amount of information. Instead, with a lower horizontal field of view, the users can turn their heads to the sides to make a lateral scan of the visual scene to temporally separate the complex scene into a succession of less complex ones. To limit the amount of auditory information transmitted simultaneously, the SSD proposed by Neugebauer et al. (2020) transmits only the central column of the image through the soundscape, which corresponds to a resolution of about 5.6°. However, the low field of view has the drawback of limiting the detection of lateral obstacles or mobile objects coming from the side, which is common in a daily use. Therefore, the size of the field of view of the camera should be determined as a trade-off between transmitting relevant information without overloading the perceptual system. In the current SSD, we incorporated the principle of reducing the amount of auditory information transmitted at one time by conveying only new visual information in the field of view of the camera, using frame differencing in the image processing step. In a real context of SSD use, many parameters other than the density of the obstacles or their relative locations may influence the abilities to use the SSD.

Reducing the flow of auditory information seems of first interest to prevent cognitive overload when other relevant information has to be perceived and processed in the scene. For instance, using an SSD to navigate in the context of a dual task while also practicing a tactile discrimination task can cause a slight increase in the time of travel when the scene configuration is complex (Stoll et al., 2015). However, the advantage of the use of spatialized sound to navigate in comparison to spatial language (i.e., “turn left”, “straight”) when the cognitive load increase has

been shown by Klatzky et al. (2006) which highlights the advantage of SSD for locomotion assistance in the context of navigation in complex environments.

In the context of an SSD conveying elevation through pitch modulation, both the vertical resolution of the processed image (e.g., 120 pixels in the current SSD) and the frequency range (e.g., [250, 1492 Hz] in the current SSD) must be considered to facilitate elevation discrimination abilities. These two parameters intrinsically defined the expectable perceived resolution through the SSD soundscape.

The current study was conducted with sighted blindfolded participants and remains to be tested with blind participants. The localization task with distractors used in the current study is comparable to a cocktail party configuration as described in Feierabend et al. (2019). They showed localization impairment in a cocktail party configuration, although the localization performance of blind people was comparable to the performance of sighted (but blindfolded) participants. Replicating the current study with blind participants might result in comparable or higher performance since studies have observed higher performances in the blind in tasks on auditory localization (Doucet et al., 2005; Voss et al., 2015) and pitch discrimination (Gougoux et al., 2004). Although Doucet et al. (2005) and Voss et al. (2015) showed strong inter-individual variability in azimuth localization abilities, they suggested that higher localization abilities in the blind resulted from a more effective use of spatial spectral cues for the azimuth dimension. However, Voss et al. (2015) showed that blind participants that were more effective in the use of spectral cues for the azimuth dimension were, on the contrary, impaired in elevation localization, which suggests a different use of spectral cues rather than superior localization abilities. Since our SSD mainly conveys elevation with pitch modulation, elevation localization performance is expected to be mainly dependent on pitch perception. While the use of pitch modulation for the elevation dimension in SSD has been found intuitive for blindfolded sighted participants (Bordeau et al., 2023; Stiles & Shimojo, 2015), the cross-modal correspondence between pitch height and spatial height has been suggested to be weaker in the blind population (Deroy et al., 2016). The protocol presented in this study has been designed to be easily replicated with blind participants.

5. Conclusion

The current study investigated the early-stage ability to localize a target in a minimalist and complex scene using a visual-to-auditory SSD using spatial acoustic cues and pitch modulation to convert the image captured by the camera into a soundscape. After a brief familiarization with the SSD principles, blindfolded participants were able to perceive the location of a target in a minimalist scene composed only of the target. In the complex scene composed of a target and distractors,

participants still succeeded in determining the location of the target, with more difficulty when no target-specific spectral signature was available.

This work suggests that the abilities to segregate a complex visual scene through an SSD soundscape are dependent on the availability of a specific spectral signature when the pitch modulation is used as an acoustic cue in the SSD encoding scheme. It highlights the necessity to consider both the frequency auditory filtering of the auditory system and the resolution of the field of sonification of the SSD to facilitate segregation abilities and limit the perceptive overload, which are necessary in the context of SSDs for pedestrian locomotion assistance.

Authors Contribution

C.B. and M.A. contributed to the conception and design of the experiment and interpreted the data. C.B. executed the study, performed data analysis and wrote the manuscript in close collaboration with M.A. F.S., C.M. J.D. provided important feedback. All authors have read and approved the manuscript and contributed substantially to it.

Acknowledgements

This research was funded by the Conseil Régional de Bourgogne Franche-Comté (2020_0335), France and the Fond Européen de Développement Régional (FEDER) (BG0027904). The authors thank the Conseil Régional de Bourgogne Franche-Comté, France and the Fond Européen de Développement Régional (FEDER) for their financial support, and the Université de Bourgogne and the Centre National de la Recherche Scientifique (CNRS) for providing administrative and infrastructural support

6. References

Abboud, S., Hanassy, S., Levy-Tzedek, S., Maidenbaum, S., & Amedi, A. (2014). EyeMusic: Introducing a "visual" colorful experience for the blind using auditory sensory substitution. *Restorative Neurology and Neuroscience, 32*(2), 247–257. https://doi.org/10.3233/RNN-130338

Ahrens, A., Lund, K. D., Marschall, M., & Dau, T. (2019). Sound source localization with varying amount of visual information in virtual reality. *PLOS ONE, 14(*3), e0214603. https://doi.org/10.1371/journal.pone.0214603

Algazi, V. R., Duda, R. O., Thompson, D. M., & Avendano, C. (2001). The CIPIC HRTF database. *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, 99–102. https://doi.org/10.1109/ASPAA.2001.969552

Ambard, M., Benezeth, Y., & P, P. (2015). Mobile video-to-audio transducer and motion detection for sensory substitution. *Frontiers in ICT, 2*. https://doi.org/10.3389/fict.2015.00020

Auvray, M., Hanneton, S., & O'Regan, J. K. (2007). Learning to perceive with a visuo-auditory substitution system: Localisation and object recognition with 'The Voice'. *Perception, 36*(3), 416–430. https://doi.org/10.1068/p5631

Best, V., van Schaik, A., & Carlile, S. (2004). Separation of concurrent broadband sound sources by human listeners. *The Journal of the Acoustical Society of America, 115*(1), 324–336. https://doi.org/10.1121/1.1632484

Bordeau, C., Scalvini, F., Migniot, C., Dubois, J., & Ambard, M. (2023). Cross-modal correspondence enhances elevation localization in visual-to-auditory sensory substitution. *Frontiers in Psychology, 14*. https://doi.org/10.3389/fpsyg.2023.1079998

Bregman, A. S. (1990). Auditory scene analysis. In *The MIT Press eBooks*. https://doi.org/10.7551/mitpress/1486.001.0001

Bronkhorst, A. W. (2000). The cocktail-party problem revisited: Early processing and selection of multi-talker speech. *Attention, Perception, & Psychophysics, 77*(5), 1465–1487. https://doi.org/10.3758/s13414-015-0882-9

Brown, D., Simpson, A. J. R., & Proulx, M. J. (2015). Auditory scene analysis and sonified visual images. Does consonance negatively impact on object formation when using complex sonified stimuli? *Frontiers in Psychology, 6*. https://doi.org/10.3389/fpsyg.2015.01522

Brown, D., Macpherson, T., & Ward, J. (2011). Seeing with sound? Exploring different characteristics of a visual-to-auditory sensory substitution device. *Perception, 40*(9), 1120–1135. https://doi.org/10.1068/p6952

Brungart, D. S., Cohen, J., Cord, M., Zion, D., & Kalluri, S. (2014). Assessment of auditory spatial awareness in complex listening environments. *The Journal of the Acoustical Society of America, 136*(4), 1808–1820. https://doi.org/10.1121/1.4893932

Brungart, D., Simpson, B., & Kordik, A. (2005). Localization in the presence of multiple simultaneous sounds. *Acta Acustica United With Acustica, 91*, 471–479.

Buchs, G., Heimler, B., & Amedi, A. (2019). The effect of irrelevant environmental noise on the performance of visual-to-auditory sensory substitution devices used by blind adults. *Multisensory Research, 32*(2), 87–109. https://doi.org/10.1163/22134808-20181327

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. https://doi.org/10.1121/1.1907229

Commère, L., & Rouat, J. (2023). Evaluation of short range depth sonifications for visual-to-auditory sensory substitution. http://arxiv.org/abs/2304.05462

Commère, L., Wood, S. U. N., and Rouat, J. (2020). Evaluation of a vision-to-audition substitution system that provides 2D WHERE information and fast user learning. Techn. Rep. https://doi.org/10.48550/arXiv.2010.09041

Deroy, O., Fasiello, I., Hayward, V., & Auvray, M. (2016). Differentiated audio-tactile correspondences in sighted and blind individuals. *Journal of Experimental Psychology: Human Perception and Performance, 42*(8), 1204–1214. https://doi.org/10.1037/xhp0000152

Doucet, M.-E., Guillemot, J.-P., Lassonde, M., Gagné, J.-P., Leclerc, C., & Lepore, F. (2005). Blind subjects process auditory spectral cues more efficiently than sighted individuals. *Experimental Brain Research, 160*(2), 194–202. https://doi.org/10.1007/s00221-004-2000-4

Elli, G. V., Benetti, S., & Collignon, O. (2014). Is There a Future for Sensory Substitution Outside Academic Laboratories? *Multisensory Research, 27*(5–6), 271–291. https://doi.org/10.1163/22134808-00002460

Eramudugolla, R., Irvine, D. R. F., McAnally, K. I., Martin, R. L., & Mattingley, J. B. (2005). Directed Attention Eliminates 'Change Deafness' in Complex Auditory Scenes. *Current Biology, 15*(12), 1108–1113. https://doi.org/10.1016/j.cub.2005.05.051

Feierabend, M., Karnath, H.-O., & Lewald, J. (2019). Auditory Space Perception in the Blind : Horizontal Sound Localization in Acoustically Simple and Complex Situations. *Perception, 48*(11), 1039–1057. https://doi.org/10.1177/0301006619872062

Gougoux, F., Lepore, F., Lassonde, M., Voss, P., Zatorre, R. J., & Belin, P. (2004). Pitch discrimination in the early blind. *Nature, 430*(6997), 309–309. https://doi.org/10.1038/430309a

Hamilton-Fletcher, G., & Chan, K. C. (2021). Auditory Scene Analysis Principles Improve Image Reconstruction Abilities of Novice Vision-to-Audio Sensory Substitution Users. *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC),* 5868–5871. https://doi.org/10.1109/EMBC46164.2021.9630296

Hamilton-Fletcher, G., Mengucci, M., & Medeiros, F. (2016). Synaestheatre : sonification of coloured objects in space, in: *Proceedings of the 2016 International Conference on Live Interfaces*, pp. 252–256. Brighton, UK.

Hanneton, S., Auvray, M., & Durette, B. (2010). The Vibe : A versatile vision-to-audition sensory substitution device. *Applied Bionics and Biomechanics, 7*(4), 269–276. https://doi.org/10.1080/11762322.2010.512734

Hawley, M. L., Litovsky, R. Y., & Culling, J. F. (2004). The benefit of binaural hearing in a cocktail party : Effect of location and type of interferer. *The Journal of the Acoustical Society of America, 115*(2), 833–843. https://doi.org/10.1121/1.1639908

Howard, P. (1966). Human Spatial Orientation.I. P. Howard, and W. B. Templeton. John Wiley, London. 1966. 533 pp. Diagrams. 84s. *The Journal of the Royal Aeronautical Society, 70*(670), 960–961. https://doi.org/10.1017/S0368393100082778

Kawashima, T., & Sato, T. (2015). Perceptual limits in a simulated "Cocktail party". *Attention, Perception, & Psychophysics, 77*(6), 2108–2120. https://doi.org/10.3758/s13414-015-0910-9

Klatzky, R. L., Marston, J. R., Giudice, N. A., Golledge, R. G., & Loomis, J. M. (2006). Cognitive load of navigating without vision when guided by virtual sound versus spatial language. *Journal of Experimental Psychology: Applied, 12*(4), 223–232. https://doi.org/10.1037/1076-898X.12.4.223

Kwak, C., & Han, W. (2020). Towards Size of Scene in Auditory Scene Analysis : A Systematic Review. *Journal of Audiology and Otology, 24*(1), 1–9. https://doi.org/10.7874/jao.2019.00248

Levy-Tzedek, S., Hanassy, S., Abboud, S., Maidenbaum, S., & Amedi, A. (2012). Fast, accurate reaching movements with a visual-to-auditory sensory substitution device. *Restorative Neurology and Neuroscience, 30*(4), 313–323. https://doi.org/10.3233/RNN-2012-110219

Lorenzi, C., Gatehouse, S., & Lever, C. (1999). Sound localization in noise in normal-hearing listeners. *The Journal of the Acoustical Society of America, 105*(3), 1810–1820. https://doi.org/10.1121/1.426719

Maidenbaum, S., Abboud, S., & Amedi, A. (2014). Sensory substitution : Closing the gap between basic research and widespread practical visual rehabilitation. *Neuroscience & Biobehavioral Reviews, 41*, 3–15. https://doi.org/10.1016/j.neubiorev.2013.11.007

Makous, J. C., & Middlebrooks, J. C. (1990). Two-dimensional sound localization by human listeners. *The Journal of the Acoustical Society of America, 87*, 2188–2200. https://doi.org/10.1121/1.399186

Mendonça, C., Campos, G., Dias, P., & Santos, J. A. (2013). Learning Auditory Space : Generalization and Long-Term Effects. *PLOS ONE, 8*(10), e77900. https://doi.org/10.1371/journal.pone.0077900

Mhaish, A., Gholamalizadeh, T., Ince, G., & Duff, D. J. (2016). Assessment of a visual to spatial-audio sensory substitution system. *2016 24th Signal Processing and Communication Application Conference (SIU)*, 245–248. https://doi.org/10.1109/SIU.2016.7495723

Neugebauer, A., Rifai, K., Getzlaff, M., & Wahl, S. (2020). Navigation aid for blind persons by visual-to-auditory sensory substitution : A pilot study. *PLOS ONE, 15*(8), e0237344. https://doi.org/10.1371/journal.pone.0237344

Oldfield, S. R., & Parker, S. P. A. (1984). Acuity of Sound Localisation : A Topography of Auditory Space. I. Normal Hearing Conditions. *Perception, 13*(5), 581–600. https://doi.org/10.1068/p130581

Pourghaemi, H., Gholamalizadeh, T.,Mhaish, A., Duff, D. J., and Ince, G. (2018). Realtime shape-based sensory substitution for object localization and recognition. *Proceedings of the 11th International Conference on Advances in Computer-Human Interactions.*

Proulx, M. J., Stoerig, P., Ludowig, E., and Knoll, I. (2008). Seeing 'where through the ears: effects of learning-by-doing and long-term sensory deprivation on localization based on image-to-sound substitution. *PLOS ONE, 3*(3), e1840. https://doi.org/10.1371/journal.pone.0001840

Richardson, M., Thar, J., Alvarez, J., Borchers, J., Ward, J., & Hamilton-Fletcher, G. (2019). How Much Spatial Information Is Lost in the Sensory Substitution Process? Comparing Visual, Tactile, and Auditory Approaches. *Perception, 48*(11), 1079–1103. https://doi.org/10.1177/0301006619873194

Scalvini, F., Bordeau, C., Ambard, M., Migniot, C., & Dubois, J. (2022). Low-Latency Human-Computer Auditory Interface Based on Real-Time Vision Analysis. ICASSP 2022 - 2022 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 36–40. https://doi.org/10.1109/ICASSP43922.2022.9747094

Stiles, N. R. B., & Shimojo, S. (2015). Auditory Sensory Substitution is Intuitive and Automatic with Texture Stimuli. *Scientific Reports, 5*(1), 15628. https://doi.org/10.1038/srep15628

Stoll, C., Palluel-Germain, R., Fristot, V., Pellerin, D., Alleysson, D., & Graff, C. (2015). Navigating from a Depth Image Converted into Sound. *Applied Bionics and Biomechanics*, 2015, 1–9. https://doi.org/10.1155/2015/543492

Voss, P., Tabry, V., & Zatorre, R. J. (2015). Trade-Off in the Sound Localization Abilities of Early Blind Individuals between the Horizontal and Vertical Planes. *The Journal of Neuroscience, 35*(15), 6051–6056. https://doi.org/10.1523/JNEUROSCI.4544-14.2015

Wenzel, E. M., Arruda, M., Kistler, D. J., & Wightman, F. L. (1993). Localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America, 94*(1), 111–123. https://doi.org/10.1121/1.407089

Zhong, X., & Yost, W. A. (2017). How many images are in an auditory scene? *The Journal of the Acoustical Society of America, 141*(4), 2882–2892. https://doi.org/10.1121/1.4981118

3.3. Synthèse

L'Étude 3 avait pour objectif d'évaluer les capacités de localisation avec le DSS dont le schéma d'encodage 3-dimensionnelle avait était validé au cours de l'Étude 1 et de l'Étude 2, mais dans un environnement plus complexe où plusieurs objets devaient être ignorés en parallèle de l'objet à localiser. Les capacités de localisation d'objets avec le DSS dans les dimensions de l'azimut et de l'élévation ont été évaluées en répliquant le protocole d'évaluation de l'Étude 1 et en utilisant le schéma d'encodage modulant l'intensité et l'enveloppe sonore proposé dans l'Étude 2.

En premier lieu, les résultats de l'Étude 3 suggèrent que la modulation de la hauteur tonale dans le schéma d'encodage en fonction de l'élévation permet de préserver les capacités de localisation dans la dimension de l'élévation lorsque des objets non-pertinents sont également présents dans la scène (donc dans le paysage sonore). Ceci a également permis de confirmer que la modulation de l'enveloppe dans le schéma d'encodage testée dans l'Étude 2 préservait les capacités de localisation de l'azimut et de l'élévation en se reposant sur des indices acoustiques spatiaux ainsi que sur la modulation de la hauteur tonale.

Par ailleurs, les résultats ont révélé une modulation des capacités à localiser un objet dans la dimension de l'azimut lorsque des objets non-pertinents étaient présents dans la scène et alignés horizontalement avec l'objet à localiser. Ceci suggère que les capacités à ségréguer une scène complexe à travers le paysage sonore du DSS dépendent de la disponibilité d'une signature spectrale spécifique aux objets composant la scène. Lorsque le signal-sur-bruit diminue et qu'il n'est pas possible de se reposer sur la signature spectrale des objets pour ségréguer le paysage sonore, une altération des capacités de localisation de l'azimut peut donc être observée.

Dans le contexte du développement de DSS vision-vers-audition pour l'aide à la locomotion et à la localisation d'obstacles, il est important de préserver les capacités de localisation d'obstacles lorsque l'environnement comprend de multiples obstacles. Les résultats de l'Étude 3 suggèrent qu'aux premiers stades de l'utilisation du DSS, la modulation de la hauteur tonale dans le schéma d'encodage présente l'avantage de préserver les capacités de ségrégation et de localisation lorsque des objets sont situés à différentes hauteurs. À l'inverse, lorsque de nombreux objets sont situés à des hauteurs similaires mais à différentes positions latérales, les capacités de ségrégation de la scène en se reposant sur le paysage sonore semblent altérées, ce qui nécessite des investigations supplémentaires pour réduire la probabilité de collision avec les obstacles dans un contexte d'utilisation réelle de DSS.

V. Discussion générale

La déficience visuelle a d'importantes répercussions sur la vie quotidienne des personnes touchées, telles que la diminution de l'autonomie et des difficultés de déplacements, ce qui augmente le risque d'accidents et d'isolement social. Avec plus de 337 millions de personnes déficientes visuelles dans le monde (Steinmetz et al., 2021 ; World Health Organization, 2019), accroître leur autonomie de déplacement pour garantir leur sécurité est donc un objectif important pour améliorer leur qualité de vie. Bien que la canne blanche, le chien d'aveugle, le GPS audio-guidé, et dans une moindre mesure les dispositifs de substitution sensorielle (DSSs), sont utilisés par les personnes non-voyantes, les déplacements urbains présentent toujours de nombreux dangers. Depuis les premiers dispositifs de substitution sensorielle vision-vers-tactile (TVSS) et vision-vers-audition (the vOICe) proposés respectivement par Bach-y-Rita et al. (1969) et Meijer (1992), de nombreux dispositifs ont vu le jour, mais plusieurs freins à leur adoption ont été rapportés. Pourtant, les DSSs représentent un moyen prometteur d'amélioration de l'inclusion sociale grâce à leur potentiel pour aider les personnes non-voyantes à gagner en autonomie en les aidant à percevoir et à interagir avec leur environnement au quotidien par le biais de la modalité auditive. La présente thèse s'est intégrée au projet 3D Sound Glasses (3DSG) visant à développer et à évaluer un dispositif de substitution sensorielle vision-vers-audition pour l'aide à la locomotion des personnes non-voyantes. Ce projet, porté par les laboratoires de recherche ImViA et LEAD, et financé par le Région Bourgogne-Franche-Comté, le Fond Européen de Développement Régional (FEDER) et l'Union National des Aveugles et Déficients Visuels (UNADEV), s'intéresse d'une part à la conception du dispositif d'un point de vue algorithmique pour le traitement de vidéos en temps réel ainsi que son implémentation matérielle (versant ImViA), et d'une autre part à l'optimisation des informations auditives transmises par le dispositif (versant LEAD).

Ayant pris place au sein du LEAD, les travaux de la présente thèse avaient donc pour objectif principal de déterminer un schéma d'encodage sonore pour le dispositif de substitution sensorielle d'aide à la locomotion, qui soit adapté aux capacités perceptives humaines, en évaluant les capacités de localisation d'objets. Un autre objectif important de la thèse consistait à proposer des protocoles de familiarisation avec le dispositif et d'évaluation des performances pour comparer plusieurs schémas d'encodage dans des environnements virtuels plus ou moins complexes. Le premier travail de cette thèse était de déterminer les indices acoustiques à utiliser dans le schéma d'encodage. Cette question a été abordée dans le premier axe de la thèse à travers les Études 1 et 2. Puisque l'usage d'un DSS dans un contexte d'aide à la locomotion se déroule bien souvent dans un environnement riche en obstacles, le travail de thèse a consisté dans un second temps à déterminer dans quelle mesure la localisation d'obstacles était possible dans un environnement complexe comprenant d'autres objets. Cette question a été abordée dans le second axe de la thèse à travers l'Étude 3.

1. Développement et évaluation d'un dispositif de substitution en environnement virtuel

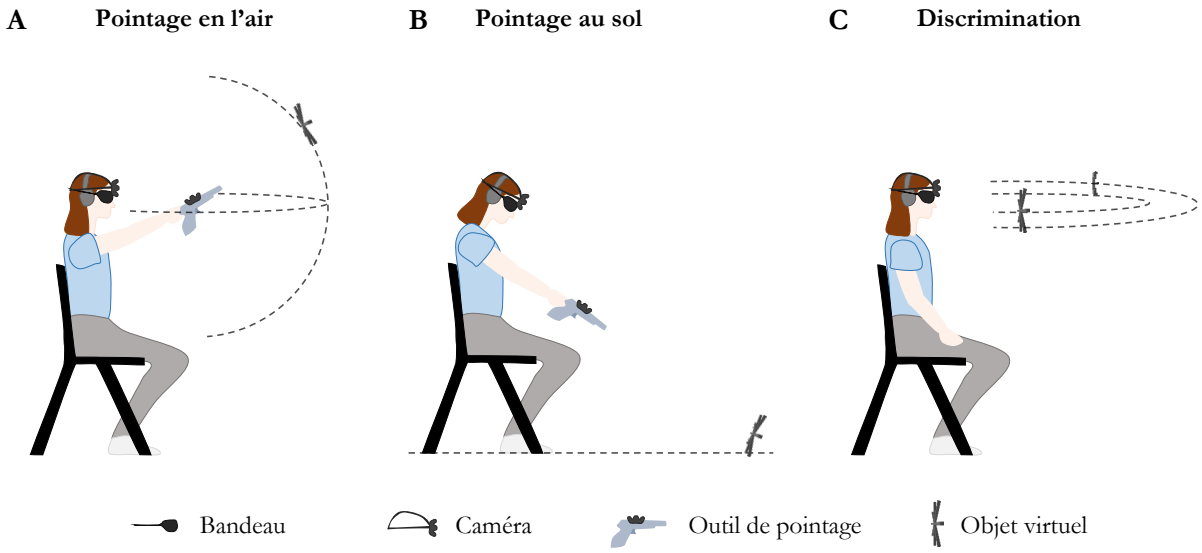
Le potentiel des environnements virtuels pour le développement et l'évaluation des capacités d'utilisation de dispositifs de substitution vision-vers-audition a été soulevé dans plusieurs revues (Elli et al., 2014 ; Kristjánsson et al., 2016 ; Real et al., 2019), et vérifié dans des études (Dascalu et al., 2017 ; Moldoveanu et al., 2017 ; Neugebauer et al., 2020 ; Real & Araujo, 2021). Les environnements virtuels permettent de mesurer précisément les capacités de perception spatiale, de standardiser les protocoles de familiarisation et d'évaluation en contrôlant les caractéristiques de l'environnement (e.g., localisation et nombre d'obstacles), ceci tout en garantissant la sécurité des participants et futurs utilisateurs (e.g., pas d'obstacles réels). Ces éléments facilitent la réplicabilité des études et permettent d'adapter la difficulté des tâches au fil de l'apprentissage. Pour toutes ces raisons, le développement des protocoles de familiarisation et l'évaluation des capacités de perception spatiale avec le dispositif ont été réalisés dans la présente thèse dans un environnement virtuel.

1.1. Deux protocoles de familiarisation audio-motrice

Pour étudier les capacités d'utilisation d'un dispositif de substitution sensorielle, la première étape était de développer des protocoles de familiarisation en environnement virtuel avec le DSS, permettant à la fois de se familiariser avec le schéma d'encodage du dispositif, mais aussi une calibration vis-à-vis de l'espace transmis par le dispositif (i.e., le champ de vision de la caméra donc le champ de sonification). Percevoir à travers un DSS, dans le sens de l'émergence d'une représentation interne de la scène environnante porteuse de sens, passe par un apprentissage perceptif sensorimoteur impliquant une recalibration spatiale (Proulx et al., 2008). Au regard de la théorie de la hiérarchie inverse (Ahissar et al., 2009), l'apprentissage perceptif passe par une recalibration des relations entre les représentations de bas niveau de traitement (les indices acoustiques utilisés dans le schéma d'encodage) et les représentations de haut niveau de traitement (des représentations sémantiques et spatiales sur la scène environnante, porteuses de sens). Pour que les sensations proximales induites par le paysage sonore puissent aboutir à des représentations de plus haut niveau et être attribuées à des objets de l'environnement externe, l'acquisition de contingences sensorimotrices est alors indispensable (Briscoe, 2018).

Dans ce cadre, deux protocoles de familiarisation ont été développés au cours de la thèse : l'un approprié spécifiquement à une familiarisation avec le schéma d'encodage de l'azimut et de l'élévation (**Figure V-1** A, utilisé dans l'Étude 1 et l'Étude 3), et l'autre spécifique à la distance (**Figure V-1**, B, utilisé dans l'Étude 2).

Figure V-1. Tâches développées pour l'évaluation des capacités de perception absolue de l'azimut et de l'élévation **(A)** et de la distance **(B)**, et pour l'évaluation des capacités de perception relative de la distance **(C)**. Les deux tâches de pointage (A et B) ont été adaptées pour être utilisées comme protocoles de familiarisation.



Premièrement, les protocoles de familiarisation ont été développés de sorte que les participants contrôlent volontairement et librement leurs actions motrices afin qu'ils puissent les associer aux modifications induites dans le paysage sonore. Dès les premiers DSSs, des études ont mis en lumière la nécessité d'actions motrices volontaires lors de la phase d'apprentissage (e.g., Auvray et al., 2005 ; Diaz, 2012 ; White et al., 1970). L'enjeu du couplage sensorimoteur est de rendre possible l'extraction des contingences sensorimotrices, c'est-à-dire des régularités entre les actions motrices et les sensations proximales auditives, comme mentionné par Auvray (2004) dans la première étape de « Contact » avec le DSS. Ce critère a été validé dans les protocoles de familiarisation de la présente thèse en permettant aux participants de placer un objet virtuel à différentes positions en dirigeant leur bras tenant l'outil de pointage dans l'Étude 1, l'Étude 2 et l'Étude 3, mais aussi en leur permettant de bouger la tête (et donc la caméra) dans l'Étude 2.

Deuxièmement, les protocoles de familiarisation ont été développés de sorte à permettre l'attribution des changements dans le paysage sonore à une modification dans l'environnement distant, c'est-à-dire à faciliter l'attribution distale (Briscoe, 2018). Si l'automatisation de l'attribution distale survient dans les derniers stades de l'apprentissage (Auvray, 2004), l'expérience phénoménologique de l'attribution distale est possible dans les premiers stades d'utilisation (Auvray, 2004 ; Auvray et al., 2005). Comme le suggèrent Auvray et al. (2005), la connaissance de la position de l'objet distal à l'origine des sensations proximales perçues avec le DSS favorise

l'inférence d'un lien de causalité entre ses sensations proximales et la source distale à l'origine de ces sensations. Dans les protocoles développés dans la présente thèse, nous avons fait en sorte que l'objet virtuel soit reconnu comme placé distalement, en l'indiquant aux participants et en leur permettant de déplacer l'objet activement et volontairement. De plus, au cours des sessions de familiarisation utilisées dans la présente thèse, les participants avaient la possibilité de déplacer l'objet virtuel en dehors du champ de vision de la caméra, résultant en une interruption de la stimulation auditive dont l'origine était l'objet virtuel. Dans l'étude de Auvray et al. (2005), l'interruption de la stimulation par la présence d'un obstacle entre les participants et l'objet à l'origine des sensations proximales facilitait l'attribution distale. Si l'interruption de la stimulation dans la présente thèse ne s'apparente pas à une obstruction de la perception telle qu'elle est conçue dans leur étude, cela a pu favoriser la constitution d'une perception spatiale distale. Dans sa revue, Briscoe (2018) affirme que le contrôle actif de la caméra est nécessaire pour le processus d'attribution distale car cela permet d'adopter un point de vue égocentrique, ce qui facilite la représentation de l'emplacement des objets à distance et la façon dont ils sont arrangés spatialement. Notons alors une différence entre d'une part le protocole de familiarisation utilisé dans l'Étude 1 et l'Étude 3, et d'autre part celui utilisé dans l'Étude 2. Dans l'Étude 2, les participants contrôlaient activement la caméra en bougeant la tête, rendant possible l'extraction des régularités sensorimotrices avec un point de vue égocentrique centré sur la caméra. À l'inverse, dans le cas de l'autre protocole de familiarisation (Étude 1 et Étude 3), les participants ne devaient pas bouger la tête, donc ne contrôlaient pas activement la caméra. Malgré cela, les résultats semblent montrer que le simple contrôle de l'objet virtuel perçu à travers le paysage sonore du DSS permette l'extraction de contingences sensorimotrices.

Ensuite, la familiarisation devait aussi permettre une recalibration spatiale pour que les participants puissent apprendre à adopter une représentation spatiale égocentrique relativement au champ de vision de la caméra (et donc au champ de sonification), comme stipulé par Proulx et al. (2008). La caméra virtuelle était positionnée sur le front des participants dans les trois études, ce qui présente l'avantage de se rapprocher d'une équivalence sensori-motrice avec le système visuel pour l'acquisition du flux vidéo à convertir (O'Regan & Noë, 2001). Le biais de sous-estimation de l'élévation observé dans l'Étude 1 avant la familiarisation tend à confirmer l'existence d'une disparité spatiale entre l'élévation de référence 0° de la caméra et l'élévation de référence propre à la représentation spatiale égocentrique des participants. En dépit de la persistance de la sous-estimation de l'élévation dans l'Étude 1 après la familiarisation, la diminution de sa magnitude suite à la familiarisation suggère qu'une recalibration spatiale a bien eu lieu grâce à la familiarisation. Néanmoins, l'absence d'un groupe contrôle n'effectuant pas de familiarisation ne permet pas d'attribuer l'origine de cette recalibration spatiale à la session de familiarisation, ou à la simple

pratique de la tâche. Néanmoins, l'Étude 3 suggère que la familiarisation y a bel et bien participé puisque la sous-estimation de l'élévation (comparable à celle de l'Étude 1) mesurée après la familiarisation ne pouvait pas être attribuée à la pratique de la tâche puisque la tâche n'était pratiquée qu'après la session de familiarisation.

La nécessité de pouvoir évaluer les DSSs auprès de la population cible non-voyante impliquait que les protocoles de familiarisation se déroulent sans nécessiter la vision. Dans la littérature, cet aspect n'est pas systématiquement considéré, rendant difficile l'utilisation de certains protocoles avec la population cible non-voyante (e.g., Ambard et al., 2015 ; Pesnot-Lerousseau et al., 2021). Pour cette raison, les protocoles de familiarisation des études présentées dans cette thèse ont été conduits auprès de participants ayant les yeux bandés. Néanmoins, mentionnons que dans l'Étude 2, durant laquelle les participants avaient le contrôle actif et volontaire de la caméra, une familiarisation les yeux ouverts précédait la familiarisation sans la vision. L'ajout de cette étape de familiarisation permettait de faciliter l'acquisition des contingences sensorimotrices et de faciliter la recalibration spatiale dans un contexte où le champ de vision de la caméra se déplaçait en temps réel avec les mouvements de la tête. Cette familiarisation les yeux ouverts permettait de s'assurer que les participants se rendent compte de l'interdépendance entre les mouvements de leur tête et les mouvements propres à la scène environnante (le déplacement de l'objet). Pour permettre ceci, mais sans la vision, cette familiarisation les yeux ouverts pourrait être remplacée par une familiarisation guidée les yeux fermés, mais avec des indications spatiales verbales de la part de l'expérimentateur (e.g., « Déplacez l'objet au sol un peu plus loin de vous » ou « Dirigez votre tête en direction du sol devant vos pieds »). Cette alternative rendrait probablement la familiarisation plus longue et augmenterait la variabilité inter-participant dans le déroulement de la familiarisation, serait effectivement implémentable avec des personnes non-voyantes.

Puisque nous souhaitions étudier les capacités à interpréter rapidement le paysage sonore du DSS aux premiers stades d'utilisation, les protocoles de familiarisation ont justement été développés pour être de courte durée. Leurs efficacités démontrées en font des options intéressantes pour être réutilisés dans des études ultérieures, d'autant plus que les entraînements de longues durées font parties des freins à l'adoption des DSSs (Hamilton-Fletcher et al., 2016b).

Pour résumer, même si les travaux de la présente thèse ne visaient pas à évaluer spécifiquement l'effet de la méthode de familiarisation avec le schéma d'encodage d'un DSS, les résultats montrent l'efficacité des deux protocoles de familiarisation qui ont été développés de sorte à répondre simultanément à cinq objectifs : **le couplage sensorimoteur** facilitant l'acquisition des contingences sensorimotrices ; **l'attribution distale** pour permettre d'inférer une origine distale aux sensations proximales auditives ; **la calibration spatiale** de l'espace égocentrique relativement

à la position et à l'angle de vue de la caméra ; **la privation visuelle** pour être adaptable à la population cible non-voyante et enfin **la courte durée**.

1.2. Trois protocoles d'évaluation des capacités de perception spatiale

Pour étudier les capacités d'utilisation d'un DSS pour l'aide à la locomotion, un autre objectif central était de déterminer des tâches expérimentales permettant d'évaluer les capacités de perception spatiale 3-dimensionnelle avec le dispositif. Tout comme pour les protocoles de familiarisation, les tâches devaient être effectuées sans avoir accès à la vision pour être adaptables à la population cible non-voyante.

De plus, pour que les tâches soient pertinentes pour évaluer les capacités de perception spatiale et qu'elles se rapprochent de situations réelles d'utilisation du DSS, le choix s'est porté sur l'évaluation des capacités de localisation d'objets, permettant d'éviter les collisions avec des obstacles lors de déplacements pédestres. Pour cela, deux tâches de localisation absolue par pointage ont été développées durant les travaux de thèse, ainsi qu'une tâche de localisation relative, toutes en environnement virtuel.

Dans la littérature, les capacités de localisation absolue avec un DSS sont souvent évaluées avec des tâches de pointage direct sur table (Auvray et al., 2007; Commère et al., 2020; Commere & Rouat, 2023; Hanneton et al., 2010; Levy-Tzedek et al., 2012; Pesnot Lerousseau et al., 2021; Pourghaemi et al., 2018; Proulx et al., 2008; Renier & De Volder, 2010), d'identification à choix forcé (Ambard et al., 2015; Brown et al., 2011; Mhaish et al., 2016; Proulx et al., 2008), de pointage direct sur écran (Bazilinskyy et al., 2016; Levy-Tzedek et al., 2012) ou de « pointage » sur une échelle analogique visuelle (Bazilinskyy et al., 2016), alors que les capacités de localisation relative sont évaluées avec des tâches de discrimination (Commere & Rouat, 2023; Richardson et al., 2019). Une synthèse des tâches utilisées est fournie dans le **Tableau II-2**.

Pour évaluer les capacités de perception absolue, les méthodes de pointage direct sur table (alternativement, attraper ou repositionner l'objet) présentent l'avantage d'être écologiques car il s'agit de tâches possibles lors de l'utilisation d'un DSS, par exemple pour attraper un parapluie sur une table. Elles permettent également une mesurabilité précise des capacités de localisation avec des métriques d'erreur ou de régression. Néanmoins, elles ne permettent pas d'étudier les capacités de localisation de l'élévation d'objet puisque les objets à localiser sont positionnés sur une même surface horizontale. Or, dans un contexte d'utilisation d'un DSS lors de déplacements urbains, il est souvent important de pouvoir détecter et distinguer les éléments en hauteur tels que des branches d'arbres ou des panneaux. Pour cette raison, nous avons adapté une tâche de localisation

basée sur des expériences de localisation auditive dans l'Étude 1 et l'Étude 3. Cette méthode utilise un outil de pointage pour localiser l'objet virtuel qui est placé à différentes positions sur une sphère centrée sur la tête de l'utilisateur (**Figure V-1** A). Utilisée dans de nombreuses tâches de localisation de sources sonores réelles et virtuelles (e.g., Makous & Middlebrooks, 1990 ; Majdak et al., 2010 ; Haber et al., 1993 ; Bahu et al., 2016), elle permet d'obtenir des métriques d'erreur et de régression pour les dimensions d'azimut et d'élévation.

Néanmoins, cette méthode de pointage, adaptée à des tâches de localisation auditive, ne permet pas d'évaluer directement les capacités de localisation de la distance d'un objet avec un DSS. Si les tâches de pointage direct sur table le permettent, elles ont la contrepartie de n'être utilisables que dans un espace restreint atteignable. Or, percevoir la distance d'un obstacle est central pour éviter les collisions. Pour cette raison, dans l'Étude 2, nous avons développé une nouvelle méthode de pointage avec un outil : le pointage au sol (**Figure V-1** B). Elle permet d'évaluer les capacités de perception de la distance d'objets avec un DSS, à la fois pour des distances proches et éloignées. Cette méthode a également permis d'obtenir à la fois des métriques d'erreur et de régression.

Au-delà des aspects de pertinence fonctionnelle et de mesurabilité précise, il y avait la nécessité que les tâches expérimentales puissent se dérouler dans des environnements virtuels plus ou moins complexes, avec une densité variée d'objets. Nous avons montré la possibilité d'adapter la tâche de pointage, utilisée à l'origine dans l'Étude 1 dans un environnement minimaliste, à un environnement plus complexe dans l'Étude 3. Comme mentionné dans la revue de Maidenbaum et al. (2014a), il est préconisé de varier progressivement la difficulté du contexte de la tâche afin de favoriser la généralisation de l'apprentissage de l'utilisation du DSS à des contextes plus complexes. La tâche développée dans l'Étude 2 pourrait aussi certainement être effectuée dans un environnement plus complexe composé d'autres objets virtuels.

En ce qui concerne les capacités de perception spatiale relative, elles ont été étudiées pour la dimension de la distance avec une tâche de discrimination adaptée à partir de celle proposée par Richardson et al. (2019) avec le dispositif Synaestheatre (**Figure V-1** C). Cette tâche nous a permis d'évaluer les capacités de discrimination de la distance de deux objets virtuels présentés simultanément. Les capacités de discrimination avec le DSS n'ont pas été évaluées pour les dimensions de l'azimut et de l'élévation au cours de la thèse. Néanmoins, un protocole d'évaluation similaire est envisageable sans difficultés majeures, comme suggéré par Richardson et al. (2019) qui ont également évalué les capacités de discrimination de l'élévation avec le dispositif.

En résumé, nous avons proposé trois tâches expérimentales répondant à quatre critères majeurs qui sont **la privation visuelle** pour être adaptables à la population cible non-voyante ; **la**

pertinence fonctionnelle en évaluant les capacités d'utilisation du dispositif dans des tâches de perception spatiale, qui sont des capacités nécessaires lors de déplacements pédestres, **la mesurabilité précise** permettant d'obtenir des métriques d'erreur et de régression, et **l'adaptabilité** permettant d'évaluer les capacités dans des environnements plus ou moins complexes.

L'importance d'évaluer les capacités d'utilisation des DSSs dans des tâches proches d'une utilisation réelle du dispositif ayant été soulevée (Elli et al., 2014), le caractère écologique des tâches développées dans cette thèse pourrait être amélioré. Par exemple, en proposant une tâche durant laquelle le participant doit rejoindre, en marchant, un objet virtuel localisé en dehors de l'espace atteignable, tel qu'utilisé lors d'expériences de localisation de la distance de sources sonores réelles (Loomis et al., 1998 ; Russell & Schneider, 2006). La mise en place d'un tel protocole est cependant plus difficile et implique une augmentation de la durée de l'expérience. Elle nécessite un grand espace et la présence d'un second expérimentateur accompagnant le participant pour garantir sa sécurité au cours des déplacements les yeux bandés. Ce protocole pourrait néanmoins être intéressant dans une future étude lors de laquelle cette tâche pourrait être couplée avec l'utilisation d'objets virtuels plus réalistes tels qu'une poubelle, un poteau, une plante ou une personne.

2. Un schéma d'encodage pour compenser les limites perceptives inhérentes aux indices acoustiques spatiaux

Le premier axe de la thèse visait à déterminer les indices acoustiques à utiliser dans le schéma d'encodage du DSS. Plus spécifiquement, il avait pour objectif d'évaluer l'avantage relatif de la reproduction d'indices acoustiques spatiaux (pour la perception spatiale auditive) et de l'utilisation d'indices acoustiques alternatifs mais impliqués dans des interactions audio-visuelles. À l'heure actuelle, il n'y a pas de consensus dans la littérature quant aux indices acoustiques à utiliser dans un DSS, mais l'importance d'une interprétation rapide a été soulignée (Hamilton-Fletcher et al., 2016b). Au regard de la théorie de l'intégration verticale dans le contexte de la perception avec un DSS, les indices acoustiques doivent être déterminés relativement à la fonctionnalité attendue et aux capacités perceptives et cognitives préexistantes des utilisateurs (Auvray et al., 2019). Ainsi, pour répondre à un besoin fonctionnel de locomotion et de localisation d'obstacles, il y avait la nécessité de se reposer sur les capacités de perception auditive spatiale pour déterminer le schéma d'encodage. De ce fait, l'intégration d'indices acoustiques spatiaux à l'aide de HRTFs semblait appropriée (détails sur le principe de spatialisation dans la section II.2.2.4). Néanmoins, la littérature s'entend à démontrer que les capacités de localisation de sources sonores simulées avec des HRTFs non-individualisées peuvent être altérées, notamment concernant la dimension de l'élévation

(détails dans la section II.2.2.2), mais aussi parce que la perception de la distance de sources sonores a tendance à être compressée (détails dans la section II.2.2.3). Certaines caractéristiques du son étant impliquées dans des correspondances cross-modales audio-visuelles ou des interactions audio-visuelles spatiales (détails dans la section II.2.4), les premiers travaux de la thèse visaient à comparer des schémas d'encodage modulant d'autres dimensions au-delà d'indices acoustiques spatiaux.

2.1. Localiser l'élévation : compenser les limites de la spatialisation avec HRTFs non-individualisées avec la correspondance audio-visuelle entre hauteur tonale et spatiale

Lors de déplacements pédestres, estimer la hauteur des obstacles est nécessaire pour différencier les obstacles posés au sol qui doivent être contournés (e.g., poubelle) des obstacles en hauteur à éviter en se baissant (e.g., branche d'arbre), et pour déterminer la hauteur des obstacles (e.g., différentiation entre un lampadaire et un petit poteau). Ceci rend nécessaire l'intégration d'indices acoustiques propres à la dimension de l'élévation dans le schéma d'encodage du DSS. Les indices acoustiques spatiaux permettant d'estimer l'élévation de sources sonores sont principalement spectraux (Blauert, 1983 pour revue, détails dans la section II.2.2.2) et il est possible de les reproduire en spatialisant un son avec des HRTFs (Li & Peissig, 2020 pour revue, détails dans la section II.2.2.4). Les capacités de localisation de l'élévation de sources sonores simulées avec des HRTFs non-individualisées étant altérées (e.g., Wenzel et al., 1993), l'utilisation d'indices acoustiques alternatifs non-spatiaux dans le schéma d'encodage du DSS paraissait pertinente, comme la hauteur tonale, qui est impliquée dans une correspondance cross-modale audio-visuelle avec l'élévation (Spence, 2011, détails dans la section II.2.4.1). La littérature scientifique n'ayant pas à ce jour identifié si les indices acoustiques spatiaux ou la modulation de la hauteur tonale était plus efficace dans le schéma d'encodage des DSSs, une étude comparative des capacités de perception de la direction des obstacles paraissait essentielle.

Les résultats de l'Étude 1 ont, d'une part, confirmé que les capacités de localisation de l'élévation de sources sonores simulées avec des HRTFs non-individualisées étaient altérées, même en utilisant un signal acoustique pouvant favoriser les performances de localisation (i.e., un bruit blanc, avec une complexité spectrale élevée et large bande). D'autre part, les résultats de l'Étude 1 ont mis en évidence un effet de facilitation de la modulation de la hauteur tonale pour localiser l'élévation d'un objet avec le DSS. Ces résultats suggèrent donc que l'utilisation de la hauteur tonale comme indice acoustique dans les DSSs permet de compenser les limites perceptives de l'élévation lorsqu'une spatialisation avec des HRTFs non-individualisées est employée. Nous attribuons cet

effet de facilitation à la correspondance cross-modale audio-visuelle entre la hauteur tonale et l'élévation.

Comme présentée dans le **Tableau II-1**, l'utilisation de la modulation de la hauteur tonale pour l'élévation dans les schémas d'encodage des DSSs est fréquente (Abboud et al., 2014 ; Ambard et al., 2015 ; Capelle et al., 1998 ; Cronly-Dillon et al., 1999 ; Hamilton-Fletcher et al., 2022 ; Hamilton-Fletcher et al., 2016a ; Hanneton et al., 2010 ; Meijer, 1992 ; Neugebauer et al., 2020 ; Stoll et al., 2015). Les travaux de la présente thèse confirment ainsi l'efficacité de cet indice acoustique pour l'élévation.

De précédents travaux ont suggéré l'aspect intuitif de cet indice acoustique pour l'élévation dans des tâches de reconnaissance d'objets avec les dispositifs the vOICe (Stiles & Shimojo, 2015) et EyeMusic (Buchs et al., 2021). L'intuitivité de cet indice acoustique pour la perception spatiale, dans une tâche de localisation, est confortée par les résultats de l'Étude 1, avec une tendance à effectivement localiser l'objet virtuel à une haute élévation lorsque le paysage sonore est plus aigu, et ce même avant la session de familiarisation avec le schéma d'encodage. Néanmoins, la qualification d'intuitivité doit être nuancée puisque les participants savaient explicitement que des caractéristiques du son étaient modifiées en fonction de la position de l'objet virtuel (bien que les caractéristiques en question n'étaient pas renseignées).

Quoi qu'il en soit, les indices spectraux contenus dans le paysage sonore propres à la modulation de la hauteur tonale étaient mieux interprétés que lorsque seulement des indices acoustiques spatiaux spectraux étaient intégrés par la spatialisation avec les HRTFs. Ceci démontre bien un effet de facilitation de la correspondance cross-modale entre la hauteur tonale et l'élévation spatiale pour localiser un objet virtuel avec un DSS. Cela a été montré dans l'Étude 1 durant laquelle les performances de localisation étaient meilleures avec la modulation de la hauteur tonale qu'avec la méthode utilisant la spatialisation d'un bruit blanc. Ce résultat va dans le sens de notre hypothèse qui prédisait des difficultés de localisation de l'élévation avec le schéma d'encodage reposant sur une combinaison de bruits blancs spatialisés, puisque les capacités de localisation de l'élévation de sources sonores spatialisées avec HRTFs non-individualisées peuvent être altérées en comparaison à des HRTFs individualisées ou des sources sonores réelles (e.g., Wenzel et al., 2013). Cependant, les performances de localisation de sources sonores spatialisées avec des HRTFs non-individualisées pouvant s'améliorer avec l'exposition ou avec l'entraînement (Berger et al., 2018 ; Mendonça, 2014 ; Stitt et al., 2019), les performances avec le DSS mériteraient d'être évaluées après une pratique prolongée ou un entraînement intensif.

Aussi, l'individualisation des HRTFs (voir Xu et al., 2007 pour revue sur les méthodes d'individualisation) pourrait être envisagée pour améliorer les performances de localisation. Si les

travaux de la présente thèse suggèrent un effet de facilitation de la hauteur tonale dans le schéma d'encodage du DSS, ils n'indiquent pas pour autant que l'unique utilisation de la spatialisation avec HRTFs pour encoder l'élévation n'est pas une méthode pertinente. Les performances de localisation de l'élévation avec ce schéma d'encodage étaient certes inférieures à celles mesurées avec le schéma d'encodage modulant la hauteur tonale, mais la perception de l'élévation s'est tout de même avérée possible. D'ailleurs, plusieurs études ont rapporté l'efficacité d'un schéma d'encodage pour l'élévation basé uniquement sur la spatialisation avec des HRTFs non-individualisées, mais dans d'autres tâches. Par exemple, Ribeiro et al. (2012) l'ont démontré dans une tâche de reconnaissance de scène par identification, et Richardson et al. (2019) l'ont démontré dans une tâche de discrimination de l'élévation en rapportant un score de discrimination de l'élévation de 14°.

De plus, l'Étude 1 a montré que la complexité spectrale et la largeur de bande des tonalités spatialisées dans les schémas d'encodage utilisant la modulation de la hauteur tonale ne modulaient pas les performances de localisation de l'élévation de l'objet. Nous nous attendions à une amélioration des performances de localisation de l'élévation lorsque des tonalités complexes étaient combinées dans le paysage sonore (condition *Harmonic*), plutôt que des tonalités pures (condition *Monotonic*). Cette hypothèse reposait sur le fait que les modifications spectrales monaurales, utilisées comme indices acoustiques spatiaux pour localiser l'élévation d'une source sonore réelle ou simulée, se concentrent principalement dans les fréquences au-delà de 4000 Hz (Algazi et al., 2001a ; Asano et al., 1990 ; Blauert, 1983 ; Gardner, 1973 ; Hebrank & Wright, 1974). Les résultats de l'Étude 1 suggèrent qu'en intégrant uniquement 2 octaves supérieures, la largeur de bande et la complexité spectrale obtenues (de 250–1000 Hz à 1492–5968 Hz) n'étaient pas suffisantes pour améliorer les capacités de localisation de l'élévation sur la base d'indices acoustiques spatiaux spectraux. Si des modifications spectrales dans des basses fréquences inférieures à 2000 Hz ont été mises en évidence (e.g., Algazi et al., 2001a), pouvant descendre jusqu'à 700 Hz (e.g., Garner, 1973, détails dans la section II.2.2.2), les stimuli auditifs utilisés dans ces études mentionnées (bruits filtrés) étaient caractérisés par une plus grande complexité spectrale que les tonalités utilisées dans le schéma d'encodage du présent DSS. De par la faible complexité spectrale et largeur de bande d'une tonalité, la localisation de l'élévation de tonalités est d'ailleurs fortement compromise (Goossens & van Opstal, 1999). Dans le cas de tonalités combinées dans le paysage sonore d'un DSS modulant la hauteur tonale pour l'élévation, il serait intéressant d'étudier si les indices acoustiques spatiaux pour l'élévation intégrés avec les HRTFs présentent véritablement un avantage, ou si la modulation de la hauteur tonale est suffisante pour localiser l'élévation. Néanmoins, spatialiser les tonalités uniquement en azimut (e.g., en fixant l'élévation à 0°) pourrait

avoir le désavantage d'altérer les capacités de localisation de l'azimut, puisque les indices spatiaux binauraux sont également modulés par l'élévation, notamment l'ILD (Schnupp et al., 2011, p. 180).

Un autre aspect qui peut être relevé dans cette étude est celui de la tendance à localiser l'élévation de l'objet virtuel à une position plus haute avec le schéma d'encodage *Harmonic* comparativement au schéma d'encodage *Monotonic* avant la familiarisation. En d'autres termes, avant que les participants n'aient connaissance du champ de sonification du DSS, l'élévation était jugée plus haute lorsque des fréquences plus aigües composaient le paysage sonore. Cette observation soulève la question de la représentation spatiale absolue ou relative de la hauteur tonale. Alors que Deroy et al. (2018) suggéraient dans leur revue que la représentation serait relative, ce résultat suggère à l'inverse qu'il pourrait être absolu, c'est-à-dire que des fréquences (ou intervalles de fréquences) seraient associées à des élévations de l'espace (ou intervalles d'élévation). Pourtant, l'étude de Pedley et Harper (1959) suggérait que la représentation spatiale de la hauteur tonale était bien relative. Dans leur étude, ils observaient qu'une même tonalité était localisée à une position plus haute lorsque les autres tonalités pouvant être diffusées durant l'expérience étaient plus graves, comparativement à lorsque les autres tonalités pouvant être diffusées durant l'expérience étaient plus aiguës. Les travaux de cette thèse soulignent ainsi l'intérêt des DSSs pour étudier les mécanismes de perception multisensorielle.

2.2. Localiser la distance : compenser la surestimation des distances proches en manipulant l'enveloppe

Lors de déplacements pédestres, estimer la distance des obstacles est essentiel pour éviter les collisions et garantir la sécurité des personnes non-voyantes, rendant indispensable l'intégration d'indices acoustiques propres à la dimension de la distance dans le schéma d'encodage du DSS. Parmi les indices acoustiques spatiaux permettant d'estimer la distance d'une source sonore, l'intensité est un indice central (Blauert, 1983 ou Zahorik, 2005 pour revue, détails dans la section II.2.2.3), faisant de sa modulation un indice acoustique couramment utilisé dans les DSSs existants (e.g., dispositif Synaestheatre, Hamilton-Fletcher et al., 2016a, détails dans le **Tableau II-1**). La perception de la distance de sources sonores est caractérisée par un biais de compression qui se traduit par une tendance à surestimer la distance des sources sonores proches et sous-estimer celle des sources sonores éloignées (Zahorik, 2005 pour revue, détails dans la section II.2.3.3). Dans le contexte des DSSs, la surestimation des distances proches pose un problème majeur en risquant d'augmenter la probabilité de collisions avec des obstacles, rendant nécessaire d'adapter le schéma d'encodage du DSS afin d'améliorer la précision dans l'estimation de la distance des objets. L'enveloppe du son étant impliquée dans la perception du timbre et de l'intensité (e.g. Schutz & Gillard, 2020, détails dans la section II.2.1.2.3), ainsi que dans des interactions audio-visuelles

spatiales (e.g., Grassi & Casco, 2009, détails dans la section II.2.4.2), une modification de l'enveloppe des tonalités composant le paysage sonore du DSS nous semblait intéressante, mais nécessitait une étude comparative des capacités de perception de la distance d'objets pour pouvoir l'évaluer.

Les résultats de l'Étude 2 ont, d'une part, confirmé un biais de compression dans la perception de la distance, et d'autre part, suggéré qu'il était possible de compenser ce biais en introduisant une modulation gaussienne de l'enveloppe des tonalités utilisées dans le schéma d'encodage du DSS. La compensation du biais de compression dans cette condition peut être attribuée à plusieurs causes. Premièrement, elle pourrait être causée par un effet de l'enveloppe gaussienne qui est caractérisée par une période d'attaque et de relâchement plus abrupte en comparaison de l'enveloppe plate du second schéma d'encodage. Les sons causés par des impacts entre des objets sont très souvent caractérisés par une période de relâchement abrupte (Schutz & Gillard, 2020) et peuvent par exemple accentuer l'effet audio-visuel d'induction de rebond (Grassi & Casco, 2009). Deuxièmement, elle pourrait être causée par une modulation de la sonie en fonction de la distance qui serait différente entre les deux schémas d'encodage. En effet, l'enveloppe d'un son pouvant moduler la sonie (e.g., Neuhoff, 1998, détails dans la section II.2.1.2.3), il est possible que la modulation de l'intensité perçue en fonction de la distance, i.e., la sonie (détails dans la section II.2.1.2.2), ait été différente avec ce schéma d'encodage à enveloppe gaussienne. Ceci pourrait également être à l'origine des meilleures capacités de discrimination de la distance qui ont été mesurées avec ce schéma d'encodage.

En dépit du fait que les travaux de la présente thèse ne permettent pas d'identifier précisément l'origine de la diminution du biais de compression, l'Étude 2 a mis en évidence qu'il était bien possible de compenser en partie ce biais à travers le schéma d'encodage du DSS, pour améliorer l'estimation de la distance des obstacles proches et améliorer la sécurité des personnes non-voyantes lors de leurs déplacements pédestres. Par exemple, si son origine est une différence de sonie, cela signifie que l'intégration d'un biais dans la fonction modulatoire de l'intensité des sons composant le paysage sonore (i.e., tonalités dans notre cas) pourrait améliorer l'estimation, ce qui peut être facilement implémenté. La modulation de l'intensité pour transmettre la distance étant utilisée dans le schéma d'encodage de nombreux DSSs (Hamilton-Fletcher et al., 2022 ; Hamilton-Fletcher, et al., 2016a ; Neugebauer et al., 2020 ; Paré et al., 2021 ; Ribeiro et al., 2012 ; Richardson et al., 2019 ; Spagnol et al., 2017 ; Stoll et al., 2015), cela pourrait s'intégrer dans de nombreux dispositifs de DSSs.

Quoi qu'il en soit, il est important de garder à l'esprit que même en intégrant volontairement un biais dans le schéma d'encodage pour la distance pour compenser ce biais de compression

d'origine, il est dans les faits presque impossible de contrôler la sonie dans le paysage sonore d'un DSS en toutes circonstances. D'une part, la sonie dépend du nombre de « pixels actifs[1] » dans l'image traitée, ce qui dépend de la taille absolue de l'objet et de sa forme, mais aussi de sa taille relative dans le champ de vision de la caméra (i.e., à la fois de sa distance, et de sa position latérale), et bien évidemment de la méthode de traitement vidéo utilisée pour obtenir l'image traitée. D'autre part, cela dépend du schéma d'encodage utilisé dans le DSS, notamment lorsqu'une modulation de la hauteur tonale est utilisée. Ici encore, ces mêmes caractéristiques, évoquées ci-dessus, qui sont propres à l'objet de la scène (ou aux objets), et à la méthode de traitement vidéo vont moduler la composition spectrale du paysage sonore, influençant potentiellement la sonie (détails dans la section II.2.1.2.2., voir courbe isosonique dans la **Figure II-6**). Pour cette raison, une compensation de l'intensité des tonalités composant le paysage sonore en fonction de leur fréquence (i.e., en fonction de la position verticale du pixel de l'image à laquelle ils sont associé) a été intégrée à l'aide de la norme ISO 226:2003. De plus, lorsque plusieurs tonalités sont diffusées simultanément, les filtres auditifs de la cochlée (Moore & Glasberg, 1983, détails dans la section II.2.1.1) peuvent induire des effets de masquage (Zwicker, 1961). Dans le cas d'un paysage sonore provenant d'un DSS modulant la hauteur tonale, la caractéristique tonotopique de la cochlée influence donc très certainement la perception de la sonie du paysage sonore, et donc l'estimation de la distance d'un objet lorsque cette dimension est encodée avec la modulation de l'intensité. C'est d'ailleurs pour cette raison que dans la tâche de discrimination de l'Étude 2, les participants étaient encouragés à bouger leur tête afin d'isoler chacun à leur tour un des deux objets dans le champ de vision de la caméra.

Les stratégies audio-motrices d'exploration de l'environnement avec le DSS n'ont pas été étudiées dans la présente thèse. Par exemple, il serait intéressant de chercher à savoir si l'isolement volontaire de chacun des objets évoluait en fonction de la difficulté de la tâche de discrimination dans l'Étude 2 (i.e., à mesure de la diminution de la distance entre les deux objets virtuels). Nous pourrions alors prédire que la stratégie d'isoler alternativement chaque objet dans le champ de vision de la caméra serait davantage nécessaire avec l'augmentation de la difficulté de la tâche. Nous pourrions aussi analyser les activités motrices des participants (e.g., mouvements de tête), pour identifier des stratégies sensorimotrices, comme l'a investigué Bermejo et al. (2015) avec le dispositif the vOICe dans une tâche de reconnaissance de formes.

Dans la présente thèse, nous nous sommes concentrés sur les premiers stades d'utilisation du DSS. Il est probable qu'une amélioration des performances de localisation de la distance puisse

[1] Le terme « pixel actif » se réfère à un pixel de l'image traitée dont le pixel auditif associé (i.e., une tonalité dans notre cas) sera transmis dans le paysage sonore du dispositif. Il est qualifié d' « actif » car il est porteur d'informations.

être mesurée après une pratique prolongée. Étant donné que la familiarité avec une source sonore peut améliorer les capacités à estimer sa distance (Coleman, 1962), nous pourrions nous attendre à une amélioration des performances de localisation de la distance d'objets avec le DSS après plus de pratique. À l'inverse, si le biais de compression de la distance est inhérent à la représentation interne spatiale, nous pourrions prédire qu'une pratique intense avec le DSS n'améliorerait pas forcément ce biais.

2.3. Localiser l'azimut avec des indices acoustiques spatiaux : une surestimation latérale et des limites de ségrégation

Lors de déplacements pédestres, estimer correctement la position latérale des obstacles est tout aussi importante que leur élévation et leur distance. Pour estimer l'azimut d'une source sonore, des indices acoustiques spatiaux binauraux sont utilisés (ILD, ITD et IPD) (Blauert, 1983 pour revue, détails dans la section II.2.2.1). Ces indices acoustiques binauraux peuvent être intégrés dans le schéma d'encodage du DSS en les approximant (e.g., le dispositif the Vibe, Hanneton et al., 2010, détails dans le **Tableau II-1**) ou en spatialisant avec des HRTFs (e.g., le dispositif See Differently, Commère et al., 2020, détails dans le **Tableau II-1**). Intégrer ces indices binauraux avec des HRTFs présente l'avantage d'intégrer la potentielle modulation de l'ILD par la position en élévation (Schnupp et al., 2011, p.180). Les capacités de localisation de l'azimut de sources sonores simulées avec des HRTFs non-individualisées sont souvent mieux préservées comparativement à l'élévation, malgré une probabilité accrue de confusions devant-derrière dues au cône de confusion (e.g., Wenzel et al., 1993), ce qui rend pertinents ces indices acoustiques spatiaux binauraux. Les DSSs utilisent souvent la modulation de la hauteur tonale pour encoder l'élévation, mais la composition spectrale de sources sonores est connue pour influencer les capacités de localisation de l'azimut (Blauert, 1983 pour revue, détails dans les sections II.2.2.1 et II.2.3.1). Il était donc nécessaire de conduire une étude pour évaluer dans quelle mesure le schéma d'encodage de l'élévation (i.e., spatialisation ou modulation de la hauteur tonale) pouvait influencer les capacités de localisation de l'azimut avec le dispositif.

Les résultats de l'Étude 1 ont confirmé que l'utilisation d'indices acoustiques spatiaux intégrés à partir de HRTFs non-individualisées était efficace pour localiser la position en azimut d'un objet avec un DSS, et ce peu importe la complexité spectrale des sons composant le paysage sonore (i.e., peu importe l'utilisation d'un bruit blanc spatialisé ou de tonalités pures ou complexes). Néanmoins, l'Étude 1 a aussi mis en évidence un pattern systématique de surestimation latérale de l'objet virtuel. Ce pattern de surestimation latérale a été observé dans des études portant sur la localisation de sources sonores (e.g., Oldfield & Parker, 1984 ; Wenzel et al., 1993, détails dans la section II.2.3.1), mais il ne semble pourtant pas systématique (e.g., Tabry et al., 2013). S'il s'agit

d'un biais perceptif dû au cône de confusion, comme le suggéraient Oldfield et Parker (1984), ce biais pourrait malheureusement poser des problèmes de sécurité dans un contexte de déplacement pédestre avec le DSS, entraînant des collisions avec des obstacles localisés en réalité à une position moins excentrique que celle perçue. Face à ce problème, plusieurs solutions peuvent être envisagées. Premièrement, nous pourrions introduire volontairement un biais dans le schéma d'encodage en associant des angles d'azimut différents entre ceux estimés à partir de l'image traitée et ceux utilisés dans la base de données des HRTFs. Pour ce faire, les tonalités utilisées dans le schéma d'encodage pourraient être spatialisées avec un couple de HRTFs associé à un azimut moins excentrique. Par exemple, pour compenser un biais de surestimation latérale de 10°, les tonalités associées à une position de l'espace en azimut de +40° pourraient être spatialisées avec un couple de HRTFs associé à un azimut moins excentrique de +30°.

Deuxièmement, les capacités de localisation ayant été évaluées aux premiers stades d'utilisation du DSS, une diminution de ce pattern de surestimation latérale pourrait être observée avec la pratique ou un entraînement plus intensif, comme observé dans des études de localisation de sources sonores spatialisées avec des HRTFs non-individualisées (Parseihian & Katz, 2012 ; Stitt et al., 2019). Comme l'indique Nardini (2021), l'apprentissage de l'utilisation d'un DSS implique une recalibration spatiale propre aux indices acoustiques utilisés dans le DSS, et cette recalibration spatiale consiste également à adopter un référentiel égocentrique propre au champ de vision de la caméra (Proulx et al., 2008). Dans l'Étude 1, nous avons bien observé une tendance à la diminution de la surestimation latérale, notamment après la familiarisation audio-motrice, qui pourrait être interprétée comme une recalibration spatiale propre aux indices acoustiques utilisés dans le DSS et vis-à-vis du champ de vision de la caméra (donc du champ de sonification). Comme expliqué dans la section V.1.1, cette étude ne permet pas d'attribuer l'origine de cette recalibration spatiale à la session de familiarisation ou bien à la simple pratique de la tâche, du fait de l'absence d'un groupe contrôle n'effectuant pas de familiarisation.

Troisièmement, les capacités de localisation de l'azimut avec le DSS doivent encore être évaluées dans un paradigme *closed-loop*, c'est-à-dire dans lequel les participants pourraient bouger leur tête, donc la caméra (comme dans l'Étude 2). Cela pourrait donner lieu à la mise en place de stratégies sensorimotrices avec des mouvements de tête latéraux pour scanner la scène horizontalement et placer l'obstacle plus frontalement, à une position autour de l'axe médian. Les capacités de localisation de sources sonores autour de l'axe médian étant souvent meilleures (e.g., Makous & Middlebrooks, 1990, détails dans la section II.2.3.1), l'estimation de la position en azimut devrait être améliorée. Cela pourrait être étudié en utilisant la même tâche de localisation *open-loop* que dans l'Étude 1 (et Étude 3), mais en la comparant avec une condition *closed-loop*. Au-delà du fait

que nous devrions mesurer de meilleures performances de localisation de l'azimut, cela permettrait d'étudier les capacités de localisation dans un contexte plus écologique d'utilisation d'un DSS.

3. Les limites de ségrégation au sein d'un paysage sonore de dispositif de substitution

La détermination des indices acoustiques utilisés dans le schéma d'encodage du DSS (Étude 1 et Étude 2) s'est faite dans un environnement virtuel minimaliste comprenant uniquement l'objet virtuel à localiser[2]. Dans des conditions réelles d'utilisation d'un DSS, l'environnement est bien plus complexe, et plusieurs obstacles sont souvent présents simultanément dans le champ de vision de la caméra, donc dans le paysage sonore. L'un des freins majeurs au déploiement des DSSs auprès de la population cible non-voyante est relatif aux difficultés d'interprétation des informations auditives du DSS, qui peuvent induire une charge cognitive élevée en contexte réel d'utilisation (Collignon et al., 2009 ; Elli et al., 2014 ; Hamilton-Flectcher et al., 2016b). À ce jour, cet aspect essentiel a finalement peu été investigué dans la littérature scientifique (détails dans la section II.3.4). Par exemple, Buchs et al. (2019) avaient montré la possibilité d'utiliser le dispositif EyeMusic pour reconnaître des objets alors que des bruits environnants étaient diffusés. Leur étude suggérait que des utilisateurs non-voyants, expérimentés avec le dispositif, parvenaient à focaliser leur attention sur les informations auditives du paysage sonore, pertinentes pour la tâche, et à ignorer les bruits environnants non-pertinents. C'est ce que Maidenbaum et al. (2014b) suggéraient aussi dans un contexte plus écologique, au sein d'un supermarché, pour reconnaître des fruits et légumes avec le même dispositif. En situation réelle d'utilisation, des informations auditives qui sont non-pertinentes peuvent aussi être contenues dans le paysage sonore du DSS. Par exemple, lorsque la position d'obstacles statiques est déjà connue (e.g., un déplacement sur une trajectoire habituelle), mais qu'il est nécessaire de localiser les nouveaux potentiels obstacles (e.g., des plots de chantier) ou des obstacles dynamiques (e.g., des personnes, des trottinettes). Mais bien souvent, les obstacles nécessitant d'être perçus avec le DSS sont présents simultanément dans le champ visuel de la caméra, ce qui implique d'être en capacité de ségréguer le paysage sonore du DSS. Le second axe de cette thèse visait alors à évaluer dans quelle mesure le schéma d'encodage permettait de préserver les capacités de localisation d'un obstacle avec le DSS dans un environnement complexe comprenant d'autres objets. Cette situation, s'apparentant au problème de *cocktail-party* (Bronkhorst, 2000 ; Cherry, 1953), implique de pouvoir ségréguer le paysage sonore du DSS. Les capacités de ségrégation et de localisation de sources sonores dans une scène auditive sont connues

[2] Mentionnons tout de même la tâche de discrimination utilisée dans l'Étude 2 durant laquelle deux objets étaient présentés simultanément.

pour être influencées par de nombreux paramètres (Bregman, 1990 pour revue, détails dans la section II.2.3.4), dont le nombre de sources sonores (e.g., Brungart et al., 2005), la complexité spectrale des éléments composant la scène auditive, et sa configuration spatiale (e.g. Kawashima & Sato, 2015).

Les résultats de l'Étude 3 ont suggéré que les capacités à ségréguer une scène complexe à travers le paysage sonore du DSS dépendaient de la présence d'une signature spectrale spécifique aux objets composant la scène. D'un côté, ils ont mis en évidence que l'utilisation de la modulation de la hauteur tonale dans le schéma d'encodage permettait de préserver les capacités de localisation pour la dimension qu'elle encode (i.e., élévation dans notre cas) malgré la présence d'objets distracteurs à ignorer. D'un autre côté, cette étude a néanmoins soulevé de potentielles limites de perception spatiale dans la dimension azimut dans le cas où la scène environnante deviendrait trop complexe, et dans laquelle la possibilité de se reposer sur la signature spectrale spécifique à chaque objet pour ségréguer le paysage sonore serait limitée. Même si des investigations supplémentaires seraient nécessaires pour confirmer cette interprétation, elle semble la plus probable au regard des capacités et limites perceptives dans le contexte de l'analyse de scènes auditives (détails dans la section II.2.3.4).

Cependant, nous pouvons noter que dans cette Étude 3, les participants n'étaient à aucun moment entraînés à localiser un objet virtuel parmi d'autres objets virtuels avant la tâche dans l'environnement complexe. En effet, le protocole de familiarisation utilisé se déroulait dans un environnement minimaliste ne contenant que l'objet virtuel à déplacer, et aucun objet supplémentaire (**Figure V-1** A). Il serait donc intéressant de répliquer l'Étude 3, mais en intégrant une deuxième session de familiarisation se déroulant dans un environnement complexe comprenant d'autres objets virtuels à ignorer. Cette familiarisation serait effectuée avant la tâche de localisation prenant place dans l'environnement complexe. Maidenbaum et al. (2014a) préconisaient dans leur revue de varier progressivement la difficulté du contexte d'utilisation afin de favoriser la généralisation de l'apprentissage à utiliser un DSS, nous pourrions alors nous attendre à de meilleures capacités de ségrégation du paysage sonore en adaptant le protocole de familiarisation à la tâche.

Afin d'investiguer davantage les capacités de ségrégation du paysage sonore du DSS composé de plusieurs objets virtuels, il pourrait être envisagé de conduire une étude avec une tâche de jugement de numérosité d'objets (« Combien d'objets distincts sont présents dans le paysage sonore ? », similaire à Zhong et al., 2017 avec des sources sonores) ou de détection d'apparition d'un objet (« Indiquez lorsqu'un nouvel objet apparaît dans le paysage sonore », similaire à Eramudugolla et al., 2005 avec des sources sonores). Avec une tâche de jugement de numérosité,

nous pourrions identifier un nombre limité d'objets pouvant être perçus distinctement dans le paysage sonore du DSS. Nous observerions probablement de meilleures capacités lorsque les objets virtuels composant la scène seraient espacés les uns des autres, et localisés à différentes élévations, donc ayant des signatures spectrales distinctes. Avec une tâche de détection d'apparition d'un objet dans le paysage sonore, nous pourrions identifier les limites en termes de rapport signal-sur-bruit. La détection d'obstacles proches étant la priorité dans le contexte des DSSs, il serait intéressant d'évaluer si les capacités à détecter un objet apparaissant dans la scène seraient meilleures lorsque l'objet apparaît plus proche que les autres objets composant la scène, en comparaison d'une distance identique ou plus éloignée.

Finalement, l'Étude 3 a plusieurs implications dans le contexte des DSSs. Premièrement, en identifiant les limites de ségrégation du paysage sonore dans le cas d'environnements complexes, cela confirme l'intérêt de réduire le flux d'informations auditives couramment suggéré (e.g., Elli et al., 2014). Pour ce faire, il peut être envisagé de réduire le champ de vision horizontal de la caméra, qui était de 90° dans la présente thèse, pour le réduire à la transmission d'informations spatiales frontales, autour de l'axe médian. Cela permettrait de ne transmettre des informations spatiales que sur la zone dans laquelle les capacités de perception auditive spatiale sont les meilleures. Surtout, cela permettrait de réduire la quantité d'informations transmises. Ce choix méthodologique a par exemple été adopté par le DSS de Neugebauer et al. (2020) qui a un champ de sonification étroit, nécessitant la mise en place de stratégies sensorimotrices pour scanner la scène et percevoir des informations latérales. Néanmoins, ne plus transmettre les informations latérales peut présenter un risque en diminuant la perception d'obstacles localisés sur les côtés, ce qui peut mener à des collisions lorsque ces obstacles sont proches ou en mouvement.

Pour cette raison, une nouvelle étude sera prochainement conduite afin de comparer les capacités à se déplacer en utilisant le DSS dans un environnement virtuel en évitant des obstacles virtuels, mais écologiques (des poteaux), en comparant deux champs de sonification du dispositif. L'étude visera en premier lieu à évaluer les capacités à se déplacer jusqu'à une source sonore réelle en évitant les poteaux virtuels (i.e., preuve de concept). En second lieu, elle visera à comparer les performances en fonction du champ de sonification horizontal du dispositif, en comparant un champ similaire à celui utilisé jusqu'à présent (i.e., 90° dans la présente thèse), avec un champ horizontal réduit se limitant aux informations frontales autour de l'axe médian. En modulant la densité d'obstacles dans l'environnement virtuel, nous nous attendons à observer un effet facilitateur de la diminution du champ horizontal de sonification lorsque l'environnement deviendra trop dense en obstacles. Les résultats descriptifs préliminaires issus d'un pré-test effectué auprès d'une personne non-voyante sont présentés en **Annexe B**. Cette étude nécessite d'être

conduite auprès d'un plus grand nombre de participants afin d'évaluer l'effet de la taille du champ horizontal de sonification sur les performances de navigation (vitesse, nombre de collisions), mais les résultats du pré-test sont encourageants.

De plus, la réduction du flux d'informations auditives peut aussi être envisagée en amont, durant le processus de traitement de la vidéo en filtrant davantage les éléments graphiques sur le flux vidéo. Une solution prometteuse est d'utiliser des algorithmes d'intelligence artificielle pour segmenter les images, ou reconnaître des objets d'intérêt. Il s'agissait d'un des objectifs de travaux réalisés au sein du laboratoire d'informatique ImViA dans le cadre du projet 3DSG dans lequel s'inscrit la présente thèse. Ces travaux se focalisaient sur la détection de personnes en temps réel dans un flux vidéo à l'aide d'un réseau de neurones à convolution. Le DSS transmettait dans le paysage sonore seulement le barycentre de la position de la personne détectée dans le champ visuel de la caméra (actes de congrès ICASSP 2022 en **Annexe C**). En appliquant ce même principe avec d'autres catégories d'objets (e.g., voitures), ce type de traitement vidéo est prometteur pour réduire le flux d'informations auditives tout en transmettant les informations spatiales pertinentes.

4. Évaluer le dispositif auprès d'une population non-voyante

Au regard de la théorie de l'intégration verticale dans le contexte de la perception avec un DSS, les indices acoustiques doivent être déterminés relativement aux capacités perceptives et cognitives préexistantes des utilisateurs (Auvray et al., 2019). Jusqu'à présent, le DSS développé dans la présente thèse n'a pas encore été évalué auprès d'une population non-voyante. Or, la cécité visuelle a un impact sur la représentation spatiale de l'espace et entraîne des mécanismes de compensation, notamment auditive (pour revue, voir Voss et al., 2010 ; Kolarik et al., 2016, Collignon et al., 2009). Ces mécanismes de compensation se manifestent parfois par des capacités supranormales, ou à l'inverse par des déficits, notamment en termes de perception spatiale (détails dans la section II.2.3.5). Or, d'après la théorie de l'intégration verticale, puisque les capacités préexistantes de perception auditive et spatiale tendent à déterminer les capacités d'utilisation d'un DSS, la variabilité inter-individuelle de ces capacités doit être prise en compte (pour revue, voir Arnold et al., 2017). Cette évaluation serait d'autant plus nécessaire que la population non-voyante est caractérisée par une forte hétérogénéité en termes de capacités auditives et spatiales.

En ce qui concerne l'estimation de l'élévation avec le DSS, d'un côté nous pourrions nous attendre à des capacités supérieures à celles des personnes voyantes rapportées dans la présente thèse, sachant les capacités supérieures de discrimination des fréquences qui ont été mesurées chez des personnes non-voyantes (e.g., Gougoux et al., 2004, détails dans la section II.2.1.2.4). D'un autre côté, cette facilitation pourrait être atténuée, sachant que Deroy et al. (2016) ont remis en

question la présence d'une correspondance cross-modale entre hauteur tonale et élévation spatiale chez cette population non-voyante. Si depuis cette étude, l'étude de Hamilton-Fletcher et al. (2020) a suggéré qu'elle était tout de même présente, cela nécessiterait d'être étudié avec des travaux supplémentaires. Néanmoins, le nombre important de DSSs qui utilisent la modulation de la hauteur tonale dans le schéma d'encodage pour l'élévation, et qui ont été évalués chez des personnes non-voyantes (e.g., Buchs et al., 2021, Jicol et al., 2021 ; Maidenbaum et al., 2014b), tend à prédire des performances de localisation comparables à celles mesurées avec le DSS développé dans ces travaux de thèse.

Concernant les capacités de localisation de l'azimut et de la distance avec le DSS, qui reposent principalement sur des indices acoustiques spatiaux (ITD, ILD et intensité), nous pourrions nous attendre à une importante variabilité inter-individuelle, comme observée par Voss et al. (2010) pour la dimension de l'azimut et par Cappagli et al. (2015) pour la distance. Néanmoins, pour la dimension de la distance, nous savons que l'absence de calibration visuo-motrice de l'espace, notamment chez les personnes non-voyantes congénitales (Collignon et al., 2009 ; Cappagli et al., 2015), pourrait prédire une altération des capacités de localisation, particulièrement pour cette dimension, notamment pour les distances éloignées. Notons tout de même que la priorité avec un DSS reste l'estimation des distances proches.

Dans un environnement complexe, composé de plusieurs objets distracteurs, similaire à l'Étude 3, nous pourrions nous attendre à des capacités de localisation comparables, voire supérieures, chez des personnes non-voyantes. Effectivement, alors que Feierabend et al. (2019) montraient un déficit pour localiser une unique source sonore, ils ne montraient pas de déficit dans une situation de *cocktail party* où les participants non-voyants devaient localiser une source sonore diffusée parmi d'autres sources sonores. Cette absence d'altération des capacités de localisation auditive dans des situations complexes chez des personnes non-voyantes peut être expliquée par une réorganisation fonctionnelle, permettant de compenser l'absence d'intégration multimodale audio-visuelle pour résoudre des situations complexes telles que dans une situation de *coktail party* (Colignon et al., 2009). Ces mécanismes de compensation expliqueraient également la meilleure utilisation des indices acoustiques spectraux pour l'élévation (Voss et al., 2015) et des indices de réverbération pour la distance (Kolarik et al., 2013) qui peut être observée chez les personnes non-voyantes. Au-delà des mécanismes de compensation, une recalibration de l'espace par l'entraînement est possible chez les personnes non-voyantes, comme le suggère l'étude de Finocchietti et al. (2017) qui montre qu'un court entraînement audio-moteur de 2 minutes peut améliorer les capacités de localisation de sources sonores. Ces résultats sont encourageants au

regard de l'applicabilité des courts protocoles de familiarisation développés dans cette thèse, qui pourraient donc s'avérer efficaces aussi chez des personnes non-voyantes.

Finalement, l'évaluation du présent DSS auprès d'une population non-voyante est la prochaine étape. Comme expliqué dans la section précédente, nous avons récemment conduit un pré-test auprès d'une personne non-voyante dans une tâche de navigation et d'évitement d'obstacles virtuels avec le DSS (résultats préliminaires descriptifs en **Annexe B**). La personne a réussi à se déplacer jusqu'à une source sonore réelle située à 25 m d'elle à plusieurs reprises , en utilisant le paysage sonore du DSS pour éviter un certain nombre de poteaux virtuels. Les résultats préliminaires peuvent être considérés comme une première preuve de concept de l'efficacité du schéma d'encodage développé dans la présente thèse. Bien évidemment, l'étude nécessite d'être conduite auprès d'un plus grand nombre de participants non-voyants. Il pourrait également être envisagé de répliquer les études conduites dans la présente thèse auprès d'une population non-voyante, ce qui a été rendu possible grâce à la validation de protocoles de familiarisation et d'évaluation auprès d'une population voyante, mais ayant les yeux bandés.

5. Conclusion générale

La présente thèse s'intégrait au projet 3DSG qui visait à développer un dispositif de substitution sensorielle vision-vers-audition pour l'aide à la locomotion des personnes non-voyantes et à la localisation d'obstacles, en prenant en compte les capacités humaines perceptives auditives et spatiales. Le premier objectif de cette thèse était de déterminer un schéma d'encodage 3-dimensionnel adapté aux capacités humaines de perception auditive spatiale, en évaluant les capacités de perception spatiale avec le DSS dans un environnement virtuel minimaliste. Dans un contexte réel d'utilisation d'un DSS pour l'aide à la locomotion, les déplacements pédestres prennent place dans des environnements complexes, riches en obstacles et en informations sensorielles. Le schéma d'encodage devait alors être déterminé pour être adapté à ces situations où la scène auditive s'enrichit. Ainsi, le deuxième objectif des travaux de thèse était d'évaluer les capacités et limites de perception spatiale avec le DSS dans un environnement virtuel plus complexe.

Le développement des deux protocoles de familiarisation avec le dispositif s'est fait au regard de cinq critères qui sont : la facilitation de l'attribution distale, la possibilité d'un couplage sensorimoteur, et d'une calibration spatiale, la privation visuelle, et la courte durée. Le développement des trois protocoles d'évaluation des capacités de localisation avec le dispositif s'est fait en vérifiant quatre critères qui sont : la pertinence fonctionnelle, l'adaptabilité, la mesurabilité précise, et la privation visuelle. De façon générale, dans les trois études conduites, le paysage sonore

issu du DSS permettait de localiser un objet virtuel en ayant les yeux fermés après une familiarisation audio-motrice de moins de 5 minutes.

À travers l'Étude 1 et l'Étude 2, les travaux de la thèse ont mis en évidence qu'il était possible de compenser les limites perceptives propres aux indices acoustiques spatiaux pour l'élévation et pour la distance, en modifiant certaines caractéristiques sonores impliquées dans des interactions audio-visuelles spatiales (hauteur tonale, et enveloppe). Plus précisément, pour la perception de l'élévation d'un objet avec le DSS, l'Étude 1 a démontré un effet de facilitation de la correspondance cross-modale audio-visuelle entre la hauteur tonale et l'élévation pour compenser les limites perceptives des indices acoustiques spatiaux intégrés avec des HRTFs non-individualisées. L'Étude 3 a mis en évidence la préservation des capacités de localisation de l'élévation avec ce schéma d'encodage dans un environnement complexe composé d'objets virtuels distracteurs. Pour la perception de la distance d'un objet avec le DSS, l'Étude 2 a, d'une part, confirmé un biais de compression de la distance (dont une surestimation des distances proches), et d'autre part, a montré qu'en modifiant l'enveloppe des tonalités spatialisées composant le paysage sonore (modulation gaussienne), le biais de compression pouvait être réduit. Enfin, pour la perception de l'azimut d'un objet avec le DSS, l'Étude 1 a démontré la pertinence de l'utilisation des indices spatiaux binauraux, intégrés à partir de fonctions HRTFs non-individualisées. Néanmoins, l'Étude 3 a permis d'identifier une limite de perception spatiale pour la dimension de l'azimut lorsque des objets virtuels distracteurs étaient présents dans la scène et alignés horizontalement avec l'objet à localiser. Ces travaux soulignent la nécessité de réduire le flux d'informations auditives pour préserver les capacités de ségrégation du paysage sonore et faciliter l'utilisabilité des dispositifs de substitution.

Ce travail de thèse a ainsi permis de proposer et d'évaluer un nouveau schéma d'encodage 3-dimensionnel pour la substitution sensorielle vision-vers-audition. Une synthèse des capacités de localisation mesurées avec le DSS développé dans la présente thèse, en fonction de la complexité de l'environnement virtuel et de la présence d'un entraînement avant la tâche, est fournie en **Annexe D**. La réflexion autour du développement du présent dispositif s'est faite autour de l'aspect fonctionnel dans des situations d'usage réel. Pour limiter le flux d'informations auditives, et ainsi la surcharge perceptive et cognitive, l'accent a été mis sur la transmission d'informations visuelles pertinentes pour les déplacements pédestres, c'est-à-dire des informations nouvelles dans le champ visuel de la caméra (i.e., du mouvement). À l'issue de ce travail de thèse, le schéma d'encodage du DSS utilise des indices acoustiques binauraux pour l'azimut, combine des indices acoustiques spatiaux et la modulation de la hauteur tonale pour l'élévation, et utilise la modulation de l'intensité pour la distance. Ainsi, le paysage sonore diffusé par le DSS est constitué d'une combinaison de

tonalités pures allant de 250 à 1492 Hz, spatialisées en azimut et en élévation, dont l'enveloppe a une forme gaussienne, et dont l'intensité est modulée en fonction de la distance. Les travaux de cette thèse ne peuvent à eux seuls déterminer si les capacités de perception spatiale avec le présent DSS doivent être attribuées à l'intuitivité du schéma d'encodage, ou à l'efficacité des protocoles de familiarisation développés. Néanmoins, les deux protocoles de familiarisation étant de courtes durées (moins de 5 minutes), l'interprétation du schéma d'encodage développé semble rapide.

La place centrale de la détermination des schémas d'encodage dans l'adoptabilité des DSSs par la population cible non-voyante soulève l'importance d'évaluer et de comparer les DSSs au-delà des preuves de concept et de faisabilité des dispositifs. Avec ces travaux de thèse, nous espérons encourager la démarche d'intégrer des comparaisons de DSSs dans les phases de développement. Effectivement, à l'échelle des possibilités qu'offre la modalité auditive, une multitude d'indices acoustiques méritent d'être étudiés. Pour cette raison, trois protocoles d'évaluation des capacités de localisation 3-dimensionnelles d'objets avec un DSS ont été développés dans un environnement virtuel, pouvant être aisément adaptés à des environnements virtuels encore plus complexes (e.g., un environnement comprenant de nombreux obstacles, même en permettant le déplacement des participants). Ainsi, ces protocoles standardisés pourraient bénéficier à d'autres structures développant des DSSs pour l'aide à la locomotion. Dans l'optique de permettre une comparaison des performances de localisation entre différents DSSs, les performances mesurées avec le DSS développé dans la présente thèse en fonction de la complexité de l'environnement virtuel et de la présence d'un entraînement avant la tâche, sont résumées en **Annexe D**.

Si les travaux de la présente thèse se sont focalisés sur la comparaison des schémas d'encodage sonore du DSS, la comparaison de l'efficacité des méthodes de traitement vidéo est tout autant nécessaire. Dans cette optique, un des objectifs des travaux conduits au sein du laboratoire ImViA (Image et Vision Artificielle), dans le cadre du projet 3DSG, était de proposer une base de données de séquences vidéo de déplacements pédestres qui puisse être utilisée pendant les phases de développement de DSSs (article soumis et en cours de révision dans le journal Data in brief en **Annexe E**). La base de données, constituée de vidéos acquises lors de déplacements pédestres en environnements réel (rues et parcs dans la ville de Dijon, France) et virtuels (modèle 3D de la célèbre place Darcy à Dijon, France, créé dans le cadre du projet 3DSG) espère d'une part être un outil permettant la comparaison d'algorithmes de traitement de vidéos, et d'autre part, être une base pour la comparaison de schémas d'encodage pour la substitution sensorielle.

Lors du développement d'un DSS, il est néanmoins important de considérer que d'une façon générale, de très nombreux paramètres ont des conséquences sur le paysage sonore généré.

Des paramètres à la fois endogènes au DSS comme la méthode de traitement de la vidéo, et à la fois exogènes, comme les caractéristiques physiques des objets de la scène, régissent la quantité d'informations auditives transmises dans le paysage sonore. Il est alors essentiel d'évaluer les capacités d'utilisation dans des conditions écologiques, en rendant possible le déplacement de la caméra en temps réel (comme dans l'Étude 2), en variant le nombre d'objets présents dans la scène (comme dans l'Étude 3), mais aussi en condition de déplacements pédestres. Lors de déplacements pédestres, il peut être nécessaire de fournir à la personne non-voyante des informations de guidage en temps réel sur la direction à prendre en parallèle de la présence d'obstacles. Dans le cadre du projet 3DSG, les travaux au sein de l'ImVia ont permis de proposer une implémentation du dispositif, qui combine la détection d'obstacles avec le schéma d'encodage utilisé dans la présente thèse avec un guidage permettant de suivre un itinéraire grâce à un signal sonore spécifique supplémentaire intégré dans le paysage sonore (actes de congrès SITIS 2022 en **Annexe F**). Dans ce contexte, ces travaux ont proposé un algorithme de recherche d'itinéraire permettant de définir le plus court chemin pour atteindre une position souhaitée à partir de marqueurs placés à différentes positions d'un bâtiment. Ces travaux se focalisaient sur les déplacements en intérieur, mais des travaux sont en cours pour étendre ce système à un guidage en extérieur en utilisant un GPS.

Les travaux de cette thèse ont été conduits dans des environnements virtuels car ils permettent de mesurer précisément les performances, de contrôler l'environnement (dont le nombre d'objets et leur localisation), d'utiliser un environnement minimaliste qui peut être complexifié progressivement, et de garantir la sécurité des participants avec l'absence d'obstacles réels. À terme, l'objectif sera d'évaluer ce dispositif dans un contexte réel d'utilisation en dehors d'un environnement virtuel, ce qui implique d'une part un environnement bien plus complexe en terme de densité d'obstacles et d'informations sensorielles (e.g., bruit ambiant), et d'autre part de passer d'un environnement virtuel à un environnement réel. En dépit du fait que les travaux de la présente thèse aient pris place dans des environnements virtuels minimalistes, la modélisation 3D d'une place dense (place Darcy de Dijon, France) réalisée dans le cadre du projet 3DSG ouvre la voie à de futures études. L'objectif sera de conduire par la suite des expérimentations dans un environnement virtuel plus réaliste et complexe, comme dans l'étude de Maidenbaum et Amedi (2019), mais toujours en garantissant la sécurité des participants. Grâce à ce modèle 3D, il est envisagé de proposer une tâche de navigation dans cet environnement virtuel immersif, permettant de moduler la densité d'obstacles statiques et mobiles et d'évaluer le potentiel du DSS dans une tâche qui se rapproche de l'utilisation réelle.

En conclusion, les travaux de cette thèse ont permis de déterminer et d'évaluer un schéma d'encodage 3-dimensionnel pour un dispositif de substitution sensorielle vision-vers-audition en

développant des protocoles de familiarisation et d'évaluation en environnement virtuel, pour comparer les capacités de localisation d'obstacles avec différents schémas d'encodage. Ils ont mis en évidence, d'une part la possibilité de compenser les limites perceptives spatiales avec l'intégration d'indices acoustiques non-spatiaux dans le schéma d'encodage, et d'autre part, la nécessité de réduire le flux d'informations auditives pour préserver les capacités de ségrégation du paysage sonore. Ces travaux suggèrent qu'au-delà d'évaluer les capacités de perception avec un dispositif de substitution, il s'avère nécessaire d'identifier les limites de ces capacités, puisque les difficultés d'utilisation dans des situations réelles ont été pointées du doigt comme un frein à l'adoption d'un tel dispositif (Hamilton-Fletcher et al., 2016b). Les protocoles de familiarisation et d'évaluation en environnement virtuel ayant été développés de sorte à être adaptés à la population déficiente visuelle, ils soulignent le potentiel des environnements virtuels pour évaluer précisément les capacités d'utilisation de ces dispositifs dans un contexte sécurisé, et espèrent encourager leur utilisation dès les premiers stades de développement et d'évaluation. Désormais, les capacités d'utilisation du dispositif développé vont pouvoir être évaluées lors d'une tâche de navigation prenant place dans un environnement virtuel sécurisé, mais écologique, et auprès de personnes non-voyantes.

Références

Abboud, S., Hanassy, S., Levy-Tzedek, S., Maidenbaum, S., & Amedi, A. (2014). EyeMusic : Introducing a "visual" colorful experience for the blind using auditory sensory substitution. *Restorative Neurology and Neuroscience, 32*(2), 247–257. https://doi.org/10.3233/RNN-130338

Ahissar, M., Nahum, M., Nelken, I., & Hochstein, S. (2009). Reverse hierarchies and sensory learning. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1515), 285–299. https://doi.org/10.1098/rstb.2008.0253

Akeroyd, M. A. (2014). An overview of the major phenomena of the localization of sound sources by normal-hearing, hearing-impaired, and aided listeners. *Trends in Hearing, 18*, 233121651456044. https://doi.org/10.1177/2331216514560442

Algazi, V. R., Avendano, C., & Duda, R. O. (2001a). Elevation localization and head-related transfer function analysis at low frequencies. *The Journal of the Acoustical Society of America, 109*(3), 1110–1122. https://doi.org/10.1121/1.1349185

Algazi, V. R., Duda, R. O., Thompson, D. M., & Avendano, C. (2001b). The CIPIC HRTF database. *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, 99–102. https://doi.org/10.1109/ASPAA.2001.969552

Ambard, M., Benezeth, Y., and Pfister, P. (2015). Mobile video-to-audio transducer and motion detection for sensory substitution. *Frontiers in ICT 2*, 20. https://doi.org/10.3389/fict.2015.00020

Arnaud, L., Gracco, V., & Ménard, L. (2018). Enhanced perception of pitch changes in speech and music in early blind adults. *Neuropsychologia, 117*, 261–270. https://doi.org/10.1016/j.neuropsychologia.2018.06.009

Arno, P., Vanlierde, A., Streel, E., Wanet-Defalque, M.-C., Sanabria-Bohorquez, S., & Veraart, C. (2001). Auditory substitution of vision : Pattern recognition by the blind. *Applied Cognitive Psychology, 15*(5), 509–519. https://doi.org/10.1002/acp.720

Arnold, G., Pesnot-Lerousseau, J., & Auvray, M. (2017). Individual differences in sensory substitution. *Multisensory Research, 30*(6), 579–600. https://doi.org/10.1163/22134808-00002561

Asano, F., Suzuki, Y., & Sone, T. (1990). Role of spectral cues in median plane localization. *The Journal of the Acoustical Society of America, 88*(1), 159–168. https://doi.org/10.1121/1.399963

Ashmead, D. H., Leroy, D., & Odom, R. D. (1990). Perception of the relative distances of nearby sound sources. *Perception & Psychophysics, 47*(4), 326–331. https://doi.org/10.3758/BF03210871

Auvray, M. (2004). *Immersion et perception spatiale. L'exemple des dispositifs de substitution sensorielle*. Paris: Thèse de doctorat: Psychologie cognitive. EHESS.

Auvray, M. (2019). Multisensory and spatial processes in sensory substitution. *Restorative Neurology and Neuroscience, 37*(6), 609–619. https://doi.org/10.3233/RNN-190950

Auvray, M., Hanneton, S., Lenay, C., & O'Regan, K. (2005). There is something out there : Distal attribution in sensory substitution, twenty years later. *Journal of Integrative Neuroscience, 04*(04), 505–521. https://doi.org/10.1142/S0219635205001002

Auvray, M., Hanneton, S., & O'Regan, J. K. (2007). Learning to perceive with a visuo-auditory substitution system : localisation and object recognition with 'the Voice'. *Perception, 36*(3), 416–430. https://doi.org/10.1068/p5631

Auvray, M., & Myin, E. (2009). Perception with compensatory devices : From sensory substitution to sensorimotor extension. *Cognitive Science, 33*(6), 1036–1058. https://doi.org/10.1111/j.1551-6709.2009.01040.x

Bach-Y-Rita, P., Collins, C. C., Saunders, F. A., White, B., & Scadden, L. (1969). Vision substitution by tactile image projection. *Nature, 221*(5184), 963–964. https://doi.org/10.1038/221963a0

Bahu, H., Carpentier, T., Noisternig, M., & Warusfel, O. (2016). Comparison of different egocentric pointing methods for 3D sound localization experiments. *Acta Acustica united with Acustica, 102*(1), 107–118. https://doi.org/10.3813/AAA.918928

Bazilinskyy, P., van Haarlem, W., Quraishi, H., Berssenbrugge, C., Binda, J., & de Winter, J. (2016). Sonifying the location of an object : A comparison of three methods. *IFAC-PapersOnLine, 49*(19), 531–536. https://doi.org/10.1016/j.ifacol.2016.10.614

Begault, D. R. (1995). 3-D sound for virtual reality and multimedia. *Computer Music Journal, 19*(4), 99. https://doi.org/10.2307/3680997

Benichoux, V., Rébillat, M., & Brette, R. (2016). On the variation of interaural time differences with frequency. *The Journal of the Acoustical Society of America, 139*(4), 1810–1821. https://doi.org/10.1121/1.4944638

Berger, C. C., Gonzalez-Franco, M., Tajadura-Jiménez, A., Florencio, D., & Zhang, Z. (2018). Generic HRTFs may be good enough in virtual reality. Improving source localization through cross-modal plasticity. *Frontiers in Neuroscience, 12*(21). https://doi.org/10.3389/fnins.2018.00021

Bermejo, F., Di Paolo, E. A., Hüg, M. X., & Arias, C. (2015). Sensorimotor strategies for recognizing geometrical shapes : A comparative study with different sensory substitution devices. *Frontiers in Psychology, 6*. https://doi.org/10.3389/fpsyg.2015.00679

Best, V., Baumgartner, R., Lavandier, M., Majdak, P., & Kopčo, N. (2020). Sound externalization : A review of recent research. *Trends in Hearing, 24*, 233121652094839. https://doi.org/10.1177/2331216520948390

Best, V., van Schaik, A., & Carlile, S. (2004). Separation of concurrent broadband sound sources by human listeners. *The Journal of the Acoustical Society of America, 115*(1), 324–336. https://doi.org/10.1121/1.1632484

Bizoń-Angov, P., Osiński, D., Wierzchoń, M., & Konieczny, J. (2021). Visual echolocation concept for the colorophone sensory substitution device using virtual reality. *Sensors, 21*(1), 237. https://doi.org/10.3390/s21010237

Blauert, J. (1983). *Spatial hearing : The psychophysics of human sound localization*. MIT Press (MA).

Block, N. (2003). Tactile sensation via spatial perception. *Trends in Cognitive Sciences, 7*(7), 285–286. https://doi.org/10.1016/S1364-6613(03)00132-3

Bregman, A. S. (1990). Auditory scene analysis. Dans *The MIT Press eBooks*. https://doi.org/10.7551/mitpress/1486.001.0001

Briscoe, R. (2018). Bodily action and distal attribution in sensory substitution. Dans *British Academy eBooks* (p. 174–187). https://doi.org/10.5871/bacad/9780197266441.003.0011

Bronkhorst, A. W. (2000). The cocktail-party problem revisited : Early processing and selection of multi-talker speech. *Attention, Perception, & Psychophysics, 77*(5), 1465–1487. https://doi.org/10.3758/s13414-015-0882-9

Bronkhorst, A. W., & Houtgast, T. (1999). Auditory distance perception in rooms. *Nature, 397*(6719), 517–520. https://doi.org/10.1038/17374

Brown, D., Simpson, A. J. R., & Proulx, M. J. (2015). Auditory scene analysis and sonified visual images. Does consonance negatively impact on object formation when using complex sonified stimuli? *Frontiers in Psychology, 6*. https://doi.org/10.3389/fpsyg.2015.01522

Brown, D., Macpherson, T., & Ward, J. (2011). Seeing with sound? Exploring different characteristics of a visual-to-auditory sensory substitution Device. *Perception, 40*(9), 1120–1135. https://doi.org/10.1068/p6952

Brungart, D. S., Cohen, J., Cord, M., Zion, D., & Kalluri, S. (2014). Assessment of auditory spatial awareness in complex listening environments. *The Journal of the Acoustical Society of America, 136*(4), 1808–1820. https://doi.org/10.1121/1.4893932

Brungart, D. S., Durlach, N. I., & Rabinowitz, W. M. (1999). Auditory localization of nearby sources. II. Localization of a broadband source. *The Journal of the Acoustical Society of America, 106*(4), 1956–1968. https://doi.org/10.1121/1.427943

Brungart, D. S., Simpson, B. D., & Kordik, A. J. (2005). Localization in the presence of multiple simultaneous sounds. *Acta Acustica united with Acustica, 91*(3), 471–479.

Buchs, G., Haimler, B., Kerem, M., Maidenbaum, S., Braun, L., & Amedi, A. (2021). A self-training program for sensory substitution devices. *PLOS ONE, 16*(4), e0250281. https://doi.org/10.1371/journal.pone.0250281

Buchs, G., Heimler, B., & Amedi, A. (2019). The effect of irrelevant environmental noise on the performance of visual-to-auditory sensory substitution devices used by blind adults. *Multisensory Research, 32*(2), 87–109. https://doi.org/10.1163/22134808-20181327

Burton, M. J., Ramke, J., Marques, A. P., Bourne, R. R. A., Congdon, N., Jones, I., Ah Tong, B. A. M., Arunga, S., Bachani, D., Bascaran, C., Bastawrous, A., Blanchet, K., Braithwaite, T., Buchan, J. C., Cairns, J., Cama, A., Chagunda, M., Chuluunkhuu, C., Cooper, A., … Faal, H. B. (2021). The Lancet Global Health Commission on Global Eye Health : Vision beyond 2020. *The Lancet Global Health, 9*(4), 489–551. https://doi.org/10.1016/S2214-109X(20)30488-5

Cappagli, G., Cocchi, E., & Gori, M. (2015). Auditory and proprioceptive spatial impairments in blind children and adults. *Developmental Science, 20*(3), e12374. https://doi.org/10.1111/desc.12374

Capelle, C., Trullemans, C., Arno, P., & Veraart, C. (1998). A real-time experimental prototype for enhancement of vision rehabilitation using auditory substitution. *IEEE Transactions on Biomedical Engineering, 45*(10), 1279–1293. https://doi.org/10.1109/10.720206

Chebat, D., Harrar, V., Kupers, R., Maidenbaum, S., Amedi, A., & Ptito, M. (2017). Sensory substitution and the neural correlates of navigation in blindness. Dans *Springer eBooks* (p. 167–200). https://doi.org/10.1007/978-3-319-54446-5_6

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. https://doi.org/10.1121/1.1907229

Coleman, P. D. (1962). Failure to localize the source distance of an unfamiliar sound. *Journal of the Acoustical Society of America*, *34*(3), 345–346. https://doi.org/10.1121/1.1928121

Coleman, P. D. (1968). Dual role of frequency spectrum in determination of auditory distance. *The Journal of the Acoustical Society of America, 44*(2), 631–632. https://doi.org/10.1121/1.1911132

Collignon, O., Voss, P., Lassonde, M., & Lepore, F. (2009). Cross-modal plasticity for the spatial processing of sounds in visually deprived subjects. *Experimental Brain Research, 192*(3), 343–358. https://doi.org/10.1007/s00221-008-1553-z

Commère, L., & Rouat, J. (2022). Sonified distance in sensory substitution does not always improve localization : Comparison with a 2D and 3D handheld device (arXiv:2204.06063). arXiv. http://arxiv.org/abs/2204.06063

Commère, L., & Rouat, J. (2023). Evaluation of short range depth sonifications for visual-to-auditory sensory substitution (arXiv:2304.05462). arXiv. http://arxiv.org/abs/2304.05462

Commère, L., Wood, S. U. N., & Rouat, J. (2020). Evaluation of a vision-to-audition substitution system that provides 2D WHERE information and fast user learning. ArXiv:2010.09041. http://arxiv.org/abs/2010.09041

Cronly-Dillon, J., Persaud, K., & Gregory, R. P. F. (1999). The perception of visual images encoded in musical form : A study in cross-modality information transfer. *Proceedings of the Royal Society of London. Series B: Biological Sciences, 266*(1436), 2427–2433. https://doi.org/10.1098/rspb.1999.0942

Dascalu, M., Moldoveanu, A., Balan, O., Lupu, R. G., Ungureanu, F., & Caraiman, S. (2017). Usability assessment of assistive technology for blind and visually impaired. 2017 *E-Health and Bioengineering Conference (EHB)*, 523–526. https://doi.org/10.1109/EHB.2017.7995476

Deroy, O., Fasiello, I., Hayward, V., & Auvray, M. (2016). Differentiated audio-tactile correspondences in sighted and blind individuals. *Journal of Experimental Psychology: Human Perception and Performance, 42*(8), 1204–1214. https://doi.org/10.1037/xhp0000152

Deroy, O., Fernández-Prieto, I., Navarra, J., & Spence, C. (2018). Unraveling the paradox of spatial pitch. Dans *Cambridge University Press eBooks* (p. 77–93). https://doi.org/10.1017/9781316651247.006

Deville, B., Bologna, G., Vinckenbosch, M., & Pun, T. (2009). See COLOR: Seeing colours with an orchestra. Dans *Springer eBooks* (p. 251–279). https://doi.org/10.1007/978-3-642-00437-7_10

Díaz, Á., Barrientos, A., Jacobs, D. M., & Travieso, D. (2012). Action-contingent vibrotactile flow facilitates the detection of ground level obstacles with a partly virtual sensory substitution device. *Human Movement Science, 31*(6), 1571–1584. https://doi.org/10.1016/j.humov.2012.05.006

Doucet, M.-E., Guillemot, J.-P., Lassonde, M., Gagné, J.-P., Leclerc, C., & Lepore, F. (2005). Blind subjects process auditory spectral cues more efficiently than sighted individuals. *Experimental Brain Research, 160*(2), 194–202. https://doi.org/10.1007/s00221-004-2000-4

Elli, G. V., Benetti, S., & Collignon, O. (2014). Is there a future for sensory substitution outside academic laboratories? *Multisensory Research, 27*(5–6), 271–291. https://doi.org/10.1163/22134808-00002460

Eramudugolla, R., Irvine, D. R. F., McAnally, K. I., Martin, R. L., & Mattingley, J. B. (2005). Directed attention eliminates 'change deafness' in complex auditory scenes. *Current Biology, 15*(12), 1108–1113. https://doi.org/10.1016/j.cub.2005.05.051

Evans, K. K., & Treisman, A. (2011). Natural cross-modal mappings between visual and auditory features. *Journal of Vision, 10*(1), 6. https://doi.org/10.1167/10.1.6

Feierabend, M., Karnath, H.-O., & Lewald, J. (2019). Auditory space perception in the blind : Horizontal sound localization in acoustically simple and complex situations. *Perception, 48*(11), 1039–1057. https://doi.org/10.1177/0301006619872062

Finocchietti, S., Cappagli, G., & Gori, M. (2017). Auditory spatial recalibration in congenital blind individuals. *Frontiers in Neuroscience, 11*, 76. https://doi.org/10.3389/fnins.2017.00076

Fryer, L., Freeman, J., & Pring, L. (2014). Touching words is not enough: How visual experience influences haptic-auditory associations in the "Bouba-Kiki" effect. *Cognition, 132*(2), 164–173. https://doi.org/10.1016/j.cognition.2014.03.015

Gardner, B., & Martin, K. (1994). HRTF Measurements of a KEMAR Dummy-Head Microphone.

Gardner, M. B. (1973). Some monaural and binaural facets of median plane localization. *The Journal of the Acoustical Society of America, 54*(6), 1489–1495. https://doi.org/10.1121/1.1914447

Geronazzo, M., Sikström, E., Kleimola, J., Avanzini, F., De Götzen, A., & Serafin, S. (2018). The impact of an accurate vertical localization with HRTFs on short explorations of immersive virtual reality scenarios. *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. https://doi.org/10.1109/ismar.2018.00034

Glasberg, B. R., & Moore, B. C. J. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research, 47*(1–2), 103–138. https://doi.org/10.1016/0378-5955(90)90170-T

Gonzalez-Mora, J. L., Rodriguez-Hernandez, A., Burunat, E., Martin, F., & Castellano, M. A. (2006). Seeing the world by hearing: Virtual Acoustic Space (VAS) a new space perception system for blind people. *2006 2nd International Conference on Information & Communication Technologies*, 837–842. https://doi.org/10.1109/ICTTA.2006.1684482

Goossens, H. H. L. M., & van Opstal, A. J. (1999). Influence of head position on the spatial representation of acoustic targets. *Journal of Neurophysiology, 81*(6), 2720–2736. https://doi.org/10.1152/jn.1999.81.6.2720

Gougoux, F., Lepore, F., Lassonde, M., Voss, P., Zatorre, R. J., & Belin, P. (2004). Pitch discrimination in the early blind. *Nature, 430*(6997), 309–309. https://doi.org/10.1038/430309a

Grassi, M., & Casco, C. (2009). Audiovisual bounce-inducing effect : Attention alone does not explain why the discs are bouncing. *Journal of Experimental Psychology: Human Perception and Performance, 35*(1), 235–243. https://doi.org/10.1037/a0013031

Guezou-Philippe, A., Huet, S., Pellerin, D., & Graff, C. (2018). Prototyping and evaluating sensory substitution devices by spatial immersion in virtual environments: *Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 596–602. https://doi.org/10.5220/0006637705960602

Haber, L., Haber, R. N., Penningroth, S., Novak, K., & Radgowski, H. (1993). Comparison of nine methods of indicating the direction to objects : Data from blind adults. *Perception, 22*(1), 35–47. https://doi.org/10.1068/p220035

Haigh, A., Brown, D. J., Meijer, P., & Proulx, M. J. (2013). How well do you see what you hear? The acuity of visual-to-auditory sensory substitution. *Frontiers in Psychology, 4*. https://doi.org/10.3389/fpsyg.2013.00330

Hamilton-Fletcher, G., Alvarez, J., Obrist, M., & Ward, J. (2022). SoundSight : A mobile sensory substitution device that sonifies colour, distance, and temperature. *Journal on Multimodal User Interfaces, 16*(1), 107–123. https://doi.org/10.1007/s12193-021-00376-w

Hamilton-Fletcher, G., & Chan, K. C. (2021). Auditory scene analysis principles improve image reconstruction abilities of novice vision-to-audio sensory substitution users. *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. https://doi.org/10.1109/embc46164.2021.9630296

Hamilton-Fletcher, G., Mengucci, M. and Medeiros, F. (2016a). Synaestheatre: sonification of coloured objects in space, in: *Proceedings of the 2016 International Conference on Live Interfaces*, pp. 252–256. Brighton, UK. https://doi.org/10.13140/RG.2.1.5053.7845

Hamilton-Fletcher, G., Obrist, M., Watten, P., Mengucci, M., & Ward, J. (2016b). « I always wanted to see the night sky » : Blind user preferences for sensory substitution devices. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems,* 2162–2174. https://doi.org/10.1145/2858036.2858241

Hamilton-Fletcher, G., Pieniak, M., Stefanczyk, M., Chan, K., & Oleszkiewicz, A. (2020). Visual experience influences associations between pitch and distance, but not pitch and height. *Journal of Vision, 20*(11), 1316. https://doi.org/10.1167/jov.20.11.1316

Hanneton, S., Auvray, M., & Durette, B. (2010). The Vibe : A versatile vision-to-audition sensory substitution device. *Applied Bionics and Biomechanics, 7*(4), 269–276. https://doi.org/10.1080/11762322.2010.512734

Hanneton, S., Herquel, P., & Auvray, M. (2015). Intermodal recoding of a video game : Learning to process signals for motion perception in a pure auditory environment: Intermodal recording of a video game. *International Journal of Adaptive Control and Signal Processing, 29*(12), 1475–1483. https://doi.org/10.1002/acs.2549

Hebrank, J., & Wright, D. (1974). Spectral cues used in the localization of sound sources on the median plane. *The Journal of the Acoustical Society of America, 56*(6), 1829–1834. https://doi.org/10.1121/1.1903520

Howard, D. M., & Angus, J. A. S. (2009). *Acoustics and psychoacoustics* (4th ed). Elsevier.

Howard, P. (1966). Human Spatial Orientation.I. P. Howard, and W. B. Templeton. John Wiley, London. 1966. 533 pp. Diagrams. 84s. *The Journal of the Royal Aeronautical Society, 70*(670), 960–961. https://doi.org/10.1017/S0368393100082778

Hurley, S., & Noë, A. (2003). Neural plasticity and consciousness. *Biology & Philosophy, 18*(1), 131–168. https://doi.org/10.1023/A:1023308401356

Jesteadt, W., Wier, C. C., & Green, D. M. (1977). Intensity discrimination as a function of frequency and sensation level. *The Journal of the Acoustical Society of America, 61*(1), 169–177. https://doi.org/10.1121/1.381278

Jicol, C., Lloyd-Esenkaya, T., Proulx, M. J., Lange-Smith, S., Scheller, M., O'Neill, E., & Petrini, K. (2020). Efficiency of sensory substitution devices alone and in combination with self-motion for spatial navigation in sighted and visually impaired. *Frontiers in Psychology, 11*, 1443. https://doi.org/10.3389/fpsyg.2020.01443

Kawashima, T., & Sato, T. (2015). Perceptual limits in a simulated "Cocktail party". *Attention, Perception, & Psychophysics, 77*(6), 2108–2120. https://doi.org/10.3758/s13414-015-0910-9

Kerber, S., & Seeber, B. U. (2012). Sound localization in noise by normal-hearing listeners and cochlear implant users. *Ear & Hearing, 33*(4), 445–457. https://doi.org/10.1097/AUD.0b013e318257607b

Kim, J.-K., & Zatorre, R. J. (2008). Generalized learning of visual-to-auditory substitution in sighted individuals. *Brain Research, 1242*, 263–275. https://doi.org/10.1016/j.brainres.2008.06.038

Klatzky, R. L., Marston, J. R., Giudice, N. A., Golledge, R. G., & Loomis, J. M. (2006). Cognitive load of navigating without vision when guided by virtual sound versus spatial language. *Journal of Experimental Psychology: Applied, 12*(4), 223–232. https://doi.org/10.1037/1076-898X.12.4.223

Klein, F., Werner, S., & Mayenfels, T. (2017). Influences of training on externalization of binaural synthesis in situations of room divergence. *Journal of The Audio Engineering Society, 65*(3), 178–187. https://doi.org/10.17743/jaes.2016.0072

Kolarik, A. J., Cirstea, S., & Pardhan, S. (2013). Evidence for enhanced discrimination of virtual auditory distance among blind listeners using level and direct-to-reverberant cues. *Experimental Brain Research, 224*(4), 623–633. https://doi.org/10.1007/s00221-012-3340-0

Kolarik, A. J., Moore, B. C. J., Zahorik, P., Cirstea, S., & Pardhan, S. (2016). Auditory distance perception in humans : A review of cues, development, neuronal bases, and effects of sensory loss. *Attention, Perception, & Psychophysics, 78*(2), 373–395. https://doi.org/10.3758/s13414-015-1015-1

Kolarik, A. J., Raman, R., Moore, B. C. J., Cirstea, S., Gopalakrishnan, S., & Pardhan, S. (2020). The accuracy of auditory spatial judgments in the visually impaired is dependent on sound source distance. *Scientific Reports, 10*(1), 7169. https://doi.org/10.1038/s41598-020-64306-8

Kopčo, N., & Shinn-Cunningham, B. (2011). Effect of stimulus spectrum on distance perception for nearby sources. *Journal of the Acoustical Society of America, 130*(3), 1530–1541. https://doi.org/10.1121/1.3613705

Kristjánsson, Á., Moldoveanu, A., Jóhannesson, Ó. I., Balan, O., Spagnol, S., Valgeirsdóttir, V. V., & Unnthorsson, R. (2016). Designing sensory-substitution devices : Principles, pitfalls and potential1. *Restorative Neurology and Neuroscience, 34*(5), 769–787. https://doi.org/10.3233/RNN-160647

Kumar, S., Forster, H. M., Bailey, P., & Griffiths, T. D. (2008). Mapping unpleasantness of sounds to their auditory representation. *The Journal of the Acoustical Society of America, 124*(6), 3810–3817. https://doi.org/10.1121/1.3006380

Kupers, R., Fumal, A., de Noordhout, A. M., Gjedde, A., Schoenen, J., & Ptito, M. (2006). Transcranial magnetic stimulation of the visual cortex induces somatotopically organized qualia in blind subjects. *Proceedings of the National Academy of Sciences, 103*(35), 13256–13260. https://doi.org/10.1073/pnas.0602925103

Kwak, C., & Han, W. (2020). Towards size of scene in auditory scene analysis : A systematic review. *Journal of Audiology and Otology, 24*(1), 1–9. https://doi.org/10.7874/jao.2019.00248

Leclère, T., Lavandier, M., & Perrin, F. (2019). On the externalization of sound sources with headphones without reference to a real source. *Journal of the Acoustical Society of America, 146*(4), 2309–2320. https://doi.org/10.1121/1.5128325

Lemaître, G., Grimault, N., & Suied, C. (2017). Acoustics and psychoacoustics of sound scenes and events. Dans *Springer eBooks* (p. 41–67). https://doi.org/10.1007/978-3-319-63450-0_3

Lessard, N., Paré, M., Lepore, F., & Lassonde, M. (1998). Early-blind human subjects localize sound sources better than sighted subjects. *Nature, 395*(6699), 278–280. https://doi.org/10.1038/26228

Levy-Tzedek, S., Hanassy, S., Abboud, S., Maidenbaum, S., & Amedi, A. (2012). Fast, accurate reaching movements with a visual-to-auditory sensory substitution device. *Restorative Neurology and Neuroscience, 30*(4), 313–323. https://doi.org/10.3233/RNN-2012-110219

Lewald, J. (2002). Vertical sound localization in blind humans. *Neuropsychologia, 40*(12), 1868–1872. https://doi.org/10.1016/S0028-3932(02)00071-4

Li, S., & Peissig, J. (2020). Measurement of head-related transfer functions : A review. *Applied Sciences, 10*(14). https://doi.org/10.3390/app10145014

Lloyd-Esenkaya, T., Lloyd-Esenkaya, V., O'Neill, E., & Proulx, M. J. (2020). Multisensory inclusive design with sensory substitution. *Cognitive Research: Principles and Implications, 5*(1), 37. https://doi.org/10.1186/s41235-020-00240-7

Loomis, J. M., Klatzky, R. L., & Giudice, N. A. (2018). sensory substitution of vision : Importance of perceptual and cognitive processing. Dans *Assistive technology for blindness and low vision* (p. 179–210).

Loomis, J. M., Klatzky, R. L., Philbeck, J. W., & Golledge, R. G. (1998). Assessing auditory distance perception using perceptually directed action. *Perception & Psychophysics, 60*(6), 966–980. https://doi.org/10.3758/BF03211932

Lorenzi, C., Gatehouse, S., & Lever, C. (1999). Sound localization in noise in normal-hearing listeners. *The Journal of the Acoustical Society of America, 105*(3), 1810–1820. https://doi.org/10.1121/1.426719

Lupu, R.-G., Mitruţ, O., Stan, A., Ungureanu, F., Kalimeri, K., & Moldoveanu, A. (2020). Cognitive and affective assessment of navigation and mobility tasks for the visually impaired via electroencephalography and behavioral signals. *Sensors, 20*(20), 5821. https://doi.org/10.3390/s20205821

Macpherson, E. A., & Middlebrooks, J. C. (2002). Listener weighting of cues for lateral angle : The duplex theory of sound localization revisited. *The Journal of the Acoustical Society of America, 111*(5), 2219–2236. https://doi.org/10.1121/1.1471898

Maidenbaum, S., Abboud, S., & Amedi, A. (2014a). Sensory substitution : Closing the gap between basic research and widespread practical visual rehabilitation. *Neuroscience & Biobehavioral Reviews, 41*, 3–15. https://doi.org/10.1016/j.neubiorev.2013.11.007

Maidenbaum, S., & Amedi, A. (2019). Standardizing visual rehabilitation using simple virtual tests. *2019 International Conference on Virtual Rehabilitation (ICVR)*, 1–8. https://doi.org/10.1109/ICVR46560.2019.8994431

Maidenbaum, S., Arbel, R., Buchs, G., Shapira, S., & Amedi, A. (2014b). Vision through other senses: Practical use of sensory substitution devices as assistive technology for visual rehabilitation. *IEEE 22nd Mediterranean Conference on Control and Automation*. https://doi.org/10.1109/med.2014.6961368

Majdak, P., Goupell, M. J., & Laback, B. (2010). 3-D localization of virtual sound sources: Effects of visual environment, pointing method, and training. *Attention, Perception, & Psychophysics, 72*(2), 454–469. https://doi.org/10.3758/APP.72.2.454

Makous, J. C., & Middlebrooks, J. C. (1990). Two-dimensional sound localization by human listeners. *The Journal of the Acoustical Society of America, 87*, 2188–2200. https://doi.org/10.1121/1.399186

Martin, V., Viaud-Delmon, I., & Warusfel, O. (2021). Effect of environment-related cues on auditory distance perception in the context of audio-only augmented reality. *Applied Sciences, 12*(1), 348. https://doi.org/10.3390/app12010348

Meijer, P. B. L. (1992). An experimental system for auditory image representations. *IEEE Transactions on Biomedical Engineering, 39*(2), 112–121. https://doi.org/10.1109/10.121642

Mendonça, C., Campos, G., Dias, P., and Santos, J. A. (2013). Learning auditory space: generalization and long-term effects. *PLOS ONE 8*, e77900. https://doi.org/10.1371/journal.pone.0077900

Mendonça, C. (2014). A review on auditory space adaptations to altered head-related cues. *Frontiers in Neuroscience, 8*. https://doi.org/10.3389/fnins.2014.00219

Mershon, D. H., & Bowers, J. N. (1979). Absolute and relative cues for the auditory perception of egocentric distance. *Perception, 8*(3), 311–322. https://doi.org/10.1068/p080311

Mershon, D. H., & King, L. E. (1975). Intensity and reverberation as factors in the auditory perception of egocentric distance. *Perception & Psychophysics, 18*(6), 409–415. https://doi.org/10.3758/BF03204113

Mhaish, A., Gholamalizadeh, T., Ince, G., & Duff, D. J. (2016). Assessment of a visual to spatial-audio sensory substitution system. *2016 24th Signal Processing and Communication Application Conference (SIU)*, 245–248. https://doi.org/10.1109/SIU.2016.7495723

Miller, G. A. (1947). Sensitivity to changes in the intensity of white noise and its relation to masking and loudness. *The Journal of the Acoustical Society of America, 19*(4), 609–619. https://doi.org/10.1121/1.1916528

Miller, J. (1991). Channel interaction and the redundant-targets effect in bimodal divided attention. *Journal of Experimental Psychology: Human Perception and Performance, 17*(1), 160–169. https://doi.org/10.1037/0096-1523.17.1.160

Mills, A. W. (1958). On the minimum audible angle. *The Journal of the Acoustical Society of America, 30*(4), 237–246. https://doi.org/10.1121/1.1909553

Mohlin, P. (2011). The just audible tonality of short exponential and gaussian pure tone bursts. *The Journal of the Acoustical Society of America, 129*(6), 3827–3836. https://doi.org/10.1121/1.3573990

Moldoveanu, A., Ivaşcu, S., Stanica, I., Dascălu, M., Lupu, R. G., Ivanica, G., Bălan, O., Caraiman, S., Ungureanu, F., Moldoveanu, F., & Morar, A. (2017). Mastering an advanced sensory substitution device for visually impaired through innovative virtual training. *2017 IEEE 7th International Conference on Consumer Electronics - Berlin (ICCE-Berlin)*, 120–125. https://doi.org/10.1109/icce-berlin.2017.8210608

Moore, B. C. J. (2008). The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people. *Journal of the Association for Research in Otolaryngology, 9*(4), 399–406. https://doi.org/10.1007/s10162-008-0143-x

Moore, B. C. J., & 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. https://doi.org/10.1121/1.389861

Nardini, M. (2021). Merging familiar and new senses to perceive and act in space. *Cognitive Processing, 22*(S1), 69–75. https://doi.org/10.1007/s10339-021-01052-3

Neugebauer, A., Rifai, K., Getzlaff, M., & Wahl, S. (2020). Navigation aid for blind persons by visual-to-auditory sensory substitution : A pilot study. *PLOS ONE, 15*(8), e0237344. https://doi.org/10.1371/journal.pone.0237344

Neuhoff, J. G. (1998). Perceptual bias for rising tones. *Nature, 395*(6698), 123–124. https://doi.org/10.1038/25862

Oldfield, S. R., & Parker, S. P. A. (1984). Acuity of sound localisation : A topography of auditory space. I. Normal hearing conditions. *Perception, 13*(5), 581–600. https://doi.org/10.1068/p130581

O'Regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences, 24*(5), 939–973. https://doi.org/10.1017/S0140525X01000115

Paré, S., Bleau, M., Djerourou, I., Malotaux, V., Kupers, R., & Ptito, M. (2021). Spatial navigation with horizontally spatialized sounds in early and late blind individuals. *PLOS ONE, 16*(2), e0247448. https://doi.org/10.1371/journal.pone.0247448

Parseihian, G., Jouffrais, C., & Katz, B. F. G. (2014). Reaching nearby sources : Comparison between real and virtual sound and visual targets. *Frontiers in Neuroscience, 8*. https://doi.org/10.3389/fnins.2014.00269

Parseihian, G., & Katz, B. F. G. (2012). Rapid head-related transfer function adaptation using a virtual auditory environment. *The Journal of the Acoustical Society of America, 131*(4), 2948–2957. https://doi.org/10.1121/1.3687448

Pascual-Leone, Á., & Hamilton, R. H. (2001). Chapter 27 The metamodal organization of the brain. Dans *Elsevier eBooks* (p. 427–445). https://doi.org/10.1016/s0079-6123(01)34028-1

Pasqualotto, A., & Esenkaya, T. (2016). Sensory substitution : The spatial updating of auditory scenes "mimics" the spatial updating of visual scenes. *Frontiers in Behavioral Neuroscience, 10*. https://doi.org/10.3389/fnbeh.2016.00079

Pedley, P. E., & Harper, R. S. (1959). Pitch and the vertical localization of sound. *The American Journal of Psychology, 72*(3), 447. https://doi.org/10.2307/1420051

Perrott, D. R. (1984). Concurrent minimum audible angle : A re-examination of the concept of auditory spatial acuity. *The Journal of the Acoustical Society of America, 75*(4), 1201–1206. https://doi.org/10.1121/1.390771

Pesnot Lerousseau, J., Arnold, G., & Auvray, M. (2021). Training-induced plasticity enables visualizing sounds with a visual-to-auditory conversion device. *Scientific Reports, 11*(1), 14762. https://doi.org/10.1038/s41598-021-94133-4

Plack, C. J., & Oxenham, A. J. (2006). *The psychophysics of pitch.* Dans *Springer eBooks* (p. 7–55). https://doi.org/10.1007/0-387-28958-5_2

Pourghaemi, H., Gholamalizadeh, T., Mhaish, A., Duff, D. J., and Ince, G. (2018). Real-time shape-based sensory substitution for object localization and recognition. *Proceedings of the 11th International Conference on Advances in Computer-Human Interactions*.

Pratt, C. C. (1930). The spatial character of high and low tones. *Journal of Experimental Psychology, 13*(3), 278–285. https://doi.org/10.1037/h0072651

Proulx, M. J., Stoerig, P., Ludowig, E., & Knoll, I. (2008). Seeing 'where' through the ears: effects of learning-by-doing and long-term sensory deprivation on localization based on image-to-sound substitution. *PLOS ONE, 3*(3), e1840. https://doi.org/10.1371/journal.pone.0001840

Rayleigh, Lord. (1907). On our perception of sound direction. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 13*(74), 214–232. https://doi.org/10.1080/14786440709463595

Real, S., & Araujo, A. (2021). VES : A mixed-reality development platform of navigation systems for blind and visually impaired. *Sensors, 21*(18), 6275. https://doi.org/10.3390/s21186275

Renier, L., Collignon, O., Poirier, C., Tranduy, D., Vanlierde, A., Bol, A., Veraart, C., & Devolder, A. (2005). Cross-modal activation of visual cortex during depth perception using auditory substitution of vision. *NeuroImage, 26*(2), 573–580. https://doi.org/10.1016/j.neuroimage.2005.01.047

Renier, L., & De Volder, A. G. (2010). Vision substitution and depth perception : Early blind subjects experience visual perspective through their ears. *Disability and Rehabilitation: Assistive Technology, 5*(3), 175–183. https://doi.org/10.3109/17483100903253936

Ribeiro, F., Florencio, D., Chou, P. A., & Zhang, Z. (2012). Auditory augmented reality : Object sonification for the visually impaired. *2012 IEEE 14th International Workshop on Multimedia Signal Processing (MMSP)*, 319–324. https://doi.org/10.1109/MMSP.2012.6343462

Richardson, M., Thar, J., Alvarez, J., Borchers, J., Ward, J., & Hamilton-Fletcher, G. (2019). How much spatial information is lost in the sensory substitution process? Comparing visual, tactile,

and auditory approaches. *Perception, 48*(11), 1079–1103. https://doi.org/10.1177/0301006619873194

Ries, D. T., Schlauch, R. S., & DiGiovanni, J. J. (2008). The role of temporal-masking patterns in the determination of subjective duration and loudness for ramped and damped sounds. *The Journal of the Acoustical Society of America, 124*(6), 3772–3783. https://doi.org/10.1121/1.2999342

Risoud, M., Hanson, J.-N., Gauvrit, F., Renard, C., Lemesre, P.-E., Bonne, N.-X., & Vincent, C. (2018). Sound source localization. *European Annals of Otorhinolaryngology, Head and Neck Diseases, 135*(4), 259–264. https://doi.org/10.1016/j.anorl.2018.04.009

Rokem, A., & Ahissar, M. (2009). Interactions of cognitive and auditory abilities in congenitally blind individuals. *Neuropsychologia, 47*(3), 843–848. https://doi.org/10.1016/j.neuropsychologia.2008.12.017

Rossing, T. D., & Houtsma, A. J. M. (1986). Effects of signal envelope on the pitch of short sinusoidal tones. *The Journal of the Acoustical Society of America, 79*(6), 1926–1933. https://doi.org/10.1121/1.393199

Rusconi, E., Kwan, B., Giordano, B., Umilta, C., & Butterworth, B. (2006). Spatial representation of pitch height : The SMARC effect. *Cognition, 99*(2), 113–129. https://doi.org/10.1016/j.cognition.2005.01.004

Russell, M. K., & Schneider, A. L. (2006). Sound source perception in a two-dimensional setting : Comparison of action and nonaction-based response tasks. *Ecological Psychology, 18*(3), 223–237. https://doi.org/10.1207/s15326969eco1803_4

Scharine, A. A., Cave, K. D., & Letowski, T. R. (2009). Auditory perception and cognitive performance. Dans *Helmet-mounted displays: Sensation, perception and cognition issues*. https://doi.org/10.13140/2.1.3160.1925

Schnupp, J., Nelken, I., & King, A. (2011). *Auditory neuroscience : Making sense of sound*. MIT Press

Schutz, M., & Gillard, J. (2020). On the generalization of tones : A detailed exploration of non-speech auditory perception stimuli. *Scientific Reports, 10*(1). https://doi.org/10.1038/s41598-020-63132-2

Sekuler, R., Sekuler, A. B., & Lau, R. (1997). Sound alters visual motion perception. *Nature, 385*(6614), 308–308. https://doi.org/10.1038/385308a0

Spagnol, S., Baldan, S., & Unnthorsson, R. (2017). Auditory depth map representations with a sensory substitution scheme based on synthetic fluid sounds. *2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*, 1–6. http://ieeexplore.ieee.org/document/8122220/

Spence, C. (2011). Crossmodal correspondences : A tutorial review. *Attention, Perception, & Psychophysics, 73*(4), 971–995. https://doi.org/10.3758/s13414-010-0073-7

Steinmetz, J. D., Bourne, R. R. A., Briant, P. S., Flaxman, S. R., Taylor, H. R. B., Jonas, J. B., Abdoli, A. A., Abrha, W. A., Abualhasan, A., Abu-Gharbieh, E. G., Adal, T. G., Afshin, A., Ahmadieh, H., Alemayehu, W., Alemzadeh, S. A. S., Alfaar, A. S., Alipour, V., Androudi, S., Arabloo, J., … Vos, T. (2021). Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020 : The Right to Sight: an analysis for the Global Burden of Disease Study. *The Lancet Global Health, 9*(2), 144–160. https://doi.org/10.1016/S2214-109X(20)30489-7

Stevens, S. S., Volkmann, J., & Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America, 8*(3), 185–190. https://doi.org/10.1121/1.1915893

Stiles, N. R. B., & Shimojo, S. (2015). Auditory sensory substitution is intuitive and automatic with texture stimuli. *Scientific Reports, 5*(1), 15628. https://doi.org/10.1038/srep15628

Stitt, P., Picinali, L., & Katz, B. F. G. (2019). Auditory accommodation to poorly matched non-individual spectral localization cues through active learning. *Scientific Reports, 9*(1), 1063. https://doi.org/10.1038/s41598-018-37873-0

Stoll, C., Palluel-Germain, R., Fristot, V., Pellerin, D., Alleysson, D., & Graff, C. (2015). Navigating from a depth image converted into sound. *Applied Bionics and Biomechanics, 2015*, 1–9. https://doi.org/10.1155/2015/543492

Tabry, V., Zatorre, R. J., & Voss, P. (2013). The influence of vision on sound localization abilities in both the horizontal and vertical planes. *Frontiers in Psychology, 4*. https://doi.org/10.3389/fpsyg.2013.00932

Vallet, G. T., Shore, D. I., & Schutz, M. (2014). Exploring the role of the amplitude envelope in duration estimation. *Perception, 43*(7), 616–630. https://doi.org/10.1068/p7656

Voss, P., Collignon, O., Lassonde, M., & Leporé, F. (2010). Adaptation to sensory loss. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*(3), 308–328. https://doi.org/10.1002/wcs.13

Voss, P., Tabry, V., & Zatorre, R. J. (2015). Trade-off in the sound localization abilities of early blind individuals between the horizontal and vertical planes. *The Journal of Neuroscience, 35*(15), 6051–6056. https://doi.org/10.1523/JNEUROSCI.4544-14.2015

Wan, C. Y., Wood, A. G., Reutens, D. C., & Wilson, S. J. (2010). Early but not late-blindness leads to enhanced auditory perception. *Neuropsychologia, 48*(1), 344–348. https://doi.org/10.1016/j.neuropsychologia.2009.08.016

Wenzel, E. M., Arruda, M., Kistler, D. J., & Wightman, F. L. (1993). Localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America, 94*(1), 111–123. https://doi.org/10.1121/1.407089

Wettschureck, R. G. (1973). The absolute difference limen of directional perception in the median plane under conditions of both, natural hearing and hearing with artificial-head-system. *Acta Acustica united with Acustica, 28*(4), 197–208.

White, B. W., Saunders, F. A., Scadden, L., Bach-Y-Rita, P., & Collins, C. C. (1970). Seeing with the skin. *Perception & Psychophysics, 7*(1), 23–27. https://doi.org/10.3758/BF03210126

Wier, C. C., Jesteadt, W., & Green, D. M. (1977). Frequency discrimination as a function of frequency and sensation level. *The Journal of the Acoustical Society of America, 61*(1), 178–184. https://doi.org/10.1121/1.381251

World Health Organization. (2019). *World report on vision*. World Health Organization. https://apps.who.int/iris/handle/10665/328717

Xu, S., Li, Z., & Salvendy, G. (2007). Individualization of head-related transfer function for three-dimensional virtual auditory display: A review. Dans *Lecture Notes in Computer Science* (p. 397–407). https://doi.org/10.1007/978-3-540-73335-5_44

Yost, W. A. (2017). Sound source localization identification accuracy : Envelope dependencies. *The Journal of the Acoustical Society of America, 142*(1), 173–185. https://doi.org/10.1121/1.4990656

Zahorik, P. (2002). Assessing auditory distance perception using virtual acoustics. *The Journal of the Acoustical Society of America, 111*(4), 1832–1846. https://doi.org/10.1121/1.1458027

Zahorik, P. (2005). Auditory distance perception in humans: A summary of past and present research. *Acta Acoustica United with Acustica, 91*, 409–420.

Zahorik, P., & Wightman, F. L. (2001). Loudness constancy with varying sound source distance. *Nature Neuroscience, 4*(1), 78–83. https://doi.org/10.1038/82931

Zhong, X., & Yost, W. A. (2017). How many images are in an auditory scene? *The Journal of the Acoustical Society of America, 141*(4), 2882–2892. https://doi.org/10.1121/1.4981118

Zwicker, E. (1961). Subdivision of the audible frequency range into critical bands (Frequenzgruppen). *The Journal of the Acoustical Society of America, 33*(2), 248–248. https://doi.org/10.1121/1.1908630

Zwiers, M. P., Van Opstal, A. J., & Cruysberg, J. R. M. (2001). A spatial hearing deficit in early-blind humans. *The Journal of Neuroscience, 21*(9), RC142. https://doi.org/10.1523/JNEUROSCI.21-09-j0002.2001

Annexes

Annexe A. Étude 1 : Article publié dans Frontiers in Psychology

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 26 January 2023
DOI 10.3389/fpsyg.2023.1079998

OPEN ACCESS

EDITED BY
Benedikt Zoefel,
UMR5549 Centre de Recherche Cerveau et Cognition (CerCo), France

REVIEWED BY
Takuya Koumura,
NTT Communication Science Laboratories, Japan
Gabriel Arnold,
Independent Researcher, Villebon-sur-Yvette, France

*CORRESPONDENCE
Camille Bordeau
bordeau.camille@gmail.com

SPECIALTY SECTION
This article was submitted to Perception Science, a section of the journal Frontiers in Psychology

RECEIVED 25 October 2022
ACCEPTED 06 January 2023
PUBLISHED 26 January 2023

CITATION
Bordeau C, Scalvini F, Migniot C, Dubois J and Ambard M (2023) Cross-modal correspondence enhances elevation localization in visual-to-auditory sensory substitution. *Front. Psychol.* 14:1079998. doi: 10.3389/fpsyg.2023.1079998

Cross-modal correspondence enhances elevation localization in visual-to-auditory sensory substitution

Camille Bordeau[1]*, Florian Scalvini[2], Cyrille Migniot[2], Julien Dubois[2] and Maxime Ambard[1]

[1]LEAD-CNRS UMR5022, Université de Bourgogne, Dijon, France, [2]ImViA EA 7535, Université de Bourgogne, Dijon, France

Introduction: Visual-to-auditory sensory substitution devices are assistive devices for the blind that convert visual images into auditory images (or soundscapes) by mapping visual features with acoustic cues. To convey spatial information with sounds, several sensory substitution devices use a Virtual Acoustic Space (VAS) using Head Related Transfer Functions (HRTFs) to synthesize natural acoustic cues used for sound localization. However, the perception of the elevation is known to be inaccurate with generic spatialization since it is based on notches in the audio spectrum that are specific to each individual. Another method used to convey elevation information is based on the audiovisual cross-modal correspondence between pitch and visual elevation. The main drawback of this second method is caused by the limitation of the ability to perceive elevation through HRTFs due to the spectral narrowband of the sounds.

Method: In this study we compared the early ability to localize objects with a visual-to-auditory sensory substitution device where elevation is either conveyed using a spatialization-based only method (Noise encoding) or using pitch-based methods with different spectral complexities (Monotonic and Harmonic encodings). Thirty eight blindfolded participants had to localize a virtual target using soundscapes before and after having been familiarized with the visual-to-auditory encodings.

Results: Participants were more accurate to localize elevation with pitch-based encodings than with the spatialization-based only method. Only slight differences in azimuth localization performance were found between the encodings.

Discussion: This study suggests the intuitiveness of a pitch-based encoding with a facilitation effect of the cross-modal correspondence when a non-individualized sound spatialization is used.

KEYWORDS

Virtual Acoustic Space, spatial hearing, sound spatialization, image-to-sound conversion, cross-modal correspondence, assistive technology, visual impairment, sound source localization

1. Introduction

Visual-to-auditory Sensory substitution devices (SSDs) are assistive tools for blind people. They convert visual information into auditory information in order to convey spatial information about the surrounding environment when vision is impaired. The visual-to-auditory conversion relies on the mapping of selected visual features with specific auditory cues. Visual information is usually acquired using a camera capturing the visual scene in front of the person. Then the scene converted into auditory information is transmitted to the user through soundscapes (or auditory images) delivered with headphones.

Various visual-to-auditory encodings are used by the existing visual-to-auditory SSDs to convey spatial information. Some of them use encoding schemes based on a Virtual Acoustic Space (VAS). A VAS consists in the simulation of a binaural acoustic signature of a virtual sound source located in a 3D space. In the context of visual-to-auditory SSDs, this is mainly used to simulate sound sources at the location of the obstacles. This simulation is achieved by spatializing the sound through the incorporation of spatial auditory cues in the original monophonic sound. Then a synthesized stereophonic signal simulating the distortions occurring while receiving the audio signal by the two ears is obtained. Among the SSDs used in localization experiments, the Synaestheatre (Hamilton-Fletcher et al., 2016a; Richardson et al., 2019), the Vibe (Hanneton et al., 2010) and the one presented by Mhaish et al. (2016) spatialize azimuth (lateral position) and elevation (vertical position). Other SSDs only spatialize the azimuth: the See differently device (Rouat et al., 2014), the one studied in Ambard et al. (2015), and the recent one presented in Scalvini et al. (2022).

The generation of a VAS is based on the reproduction of binaural acoustic cues related to the relative sound source location such as timing, intensity and spectral features (for an in-depth explanation of the auditory localization mechanisms see Blauert, 1996). Those features arise from audio signal distortions mainly caused by the reflection and absorption of the head, pinna and torso and are partly reproducible using Head-Related Transfer Functions (HRTFs). HRTFs are transfer functions characterizing these signal distortions as a function of the position of the sound source relatively to the two ears. They are usually obtained by conducting multiple binaural recordings with a sound source carefully placed in various positions while repeatedly producing the same sound.

Due to the technical difficulty in acquiring these recordings in good conditions, non-individualized HRTFs acquired in controlled conditions with another listener or a manikin are frequently used. However, these HRTFs failed to simulate the variability of individual-specific spectrum distortions that are related to individual morphologies (head, torso and pinna). Consequently, the localization of simulated sound sources using non-individualized HRTFs is often inaccurate with front-back and up-down confusions that are less resolvable (Wenzel et al., 1993), and a less perceptible externalization (Best et al., 2020). Nonetheless, due to the robustness of the binaural cues, azimuth localization accuracy is well preserved compared to the perception of elevation since azimuth perception relies less on the individual-specific spectrum distortions (Makous and Middlebrooks, 1990; Wenzel et al., 1993; Middlebrooks, 1999). Therefore, visual-to-auditory encodings only based on the creation of a VAS have the advantage to rely on acoustic cues that mimic natural acoustic features for sound source localization, nevertheless in practice the elevation perception can be impaired.

To compensate for this difficulty some visual-to-auditory SSDs use additional acoustic cues to convey spatial information. For instance, pitch modulation is often used to convey elevation location (Meijer, 1992; Abboud et al., 2014; Ambard et al., 2015). This mapping between elevation location and auditory pitch is based on the audiovisual cross-modal correspondence between pitch and elevation (see Spence, 2011 for a review on audiovisual cross-modal correspondences). Humans show a tendency to associate high pitch with high spatial locations and low pitch with low spatial locations. For example, they tend to exhibit faster response times in an audio-visual Go/No-Go task when the visual and auditory stimuli are congruent, i.e., higher pitch with higher visual location, and lower pitch with lower visual location (Miller, 1991). They also tend to discriminate more accurately and quickly the location of a visual stimulus (high *vs.* low location) when the pitch of a presented sound is congruent with the visual elevation (Evans and Treisman, 2011). Also, humans tend to respond to high pitch sounds with a high-located response button instead of a lower-located response button (Rusconi et al., 2006). The pitch-based encoding used in the vOICe SSD (Meijer, 1992) has been suggested somewhat intuitive in a recognition task (Stiles and Shimojo, 2015). Nevertheless, the main drawback of a pitch-based encoding is caused by the limitation of the abilities to perceive elevation through HRTFs due to the audio spectral narrowband (Algazi et al., 2001b). Although some acoustic cues for elevation perception are present in low frequencies below 3,500 Hz (Gardner, 1973; Asano et al., 1990), localization abilities are higher when the spectral content contains high frequencies above 4,000 Hz (Hebrank and Wright, 1974; Middlebrooks and Green, 1990). Since the ability to perceive the elevation through HRTFs is higher with broadband sounds containing high frequencies, the spectral content of the sound used in the visual-to-auditory encoding might modulate the perception of elevation through HRTFs. No study has directly compared encodings only based on HRTFs with encodings adding a pitch modulation and it remains unclear if the simulation of natural acoustic cues is less efficient for object localization than a more artificial sonification method using the cross-modal correspondence between pitch and elevation.

Many studies investigating static object localization abilities have already been conducted with blindfolded sighted persons using visual-to-auditory SSDs. Various types of tasks have already been used, for example discrimination tasks with forced choice (Proulx et al., 2008; Levy-Tzedek et al., 2012; Ambard et al., 2015; Mhaish et al., 2016; Richardson et al., 2019), grasping tasks (Proulx et al., 2008), index or tool pointing tasks (Auvray et al., 2007; Hanneton et al., 2010; Brown et al., 2011; Pourghaemi et al., 2018; Commère et al., 2020), or head-pointing tasks (Scalvini et al., 2022). Those studies showed the high potential of SSDs to localize an object and interact with it. However, long trainings were often conducted before the localization tasks to learn the visual-to-auditory encoding schemes: from 5 min in Pourghaemi et al. (2018) to 3 h in Auvray et al. (2007). On the contrary, in the study of Scalvini et al. (2022) the experimenter only explained verbally the encoding schemes to the participants.

Virtual environments are more and more used to investigate the abilities to perceive the environment with a visual-to-auditory SSD (Maidenbaum et al., 2014; Kristjánsson et al., 2016) since they allow a complete control of the experimental environment (e.g., number of objects, object locations...) (Maidenbaum and Amedi, 2019) and a more accurate assessment of localization abilities with precise pointing methods. They have been used in standardization tests to compare the abilities to interpret information provided by SSDs in navigation or localization tasks (Caraiman et al., 2017; Richardson et al., 2019; Jicol et al., 2020; Real and Araujo, 2021).

The current study aimed at investigating the intuitiveness of different types of visual-to-auditory encodings for the elevation in the context of object localization with a SSD. Therefore, we conducted a localization task in a virtual environment with blindfolded

participants testing a spatialization-based encoding and a pitch-based encoding. This study also aimed at assessing whether a higher spectral complexity of the sound used in a pitch-based encoding could improve the localization performance. Therefore, 2 types of pitch-based encodings were investigated: one monotonic and one harmonic with 3 octaves. We measured the localization performance for the azimuth and for the elevation. For each of these measures, we studied the effect of the visual-to-auditory encoding before and after an audio-motor familiarization of short duration.

Since the audio spatialization method was not based on individualized HRTFs, and since the pitch-based encodings were not explained to the participants, localization performance for the elevation was expected to be impaired. However, a facilitation effect of the pitch-based encodings for the elevation localization accuracy was hypothesized. Among the two pitch-based encodings, a higher elevation localization accuracy was predicted with the harmonic encoding since the sound has a higher spectral complexity. Also, the intuitiveness of the azimuth perception for all the encodings was hypothesized since it is based on less individual-specific acoustic spatial cues than elevation perception.

2. Method

2.1. Participants

Thirty eight participants were divided into two groups: the Monotonic group (19, age: M = 25.5, SD = 3.04, 6 female, 19 right-handed) and the Harmonic group (19, age: M = 24.4, SD = 3.27, 10 female, 18 right-handed). No participant reported impairments of hearing or any history of psychiatric illness or neurological disorder. The experimental protocol was approved by the local ethical committee Comité d'Ethique pour la Recherche de Université Bourgogne Franche-Comté (CERUBFC-2021-12-21-050) and followed the ethical guidelines of the Declaration of Helsinki. Written informed consent was obtained from all the participants before the experiment. No monetary compensation was given to the participants.

2.2. Visual-to-auditory conversion in the virtual environment

The visual-to-auditory SSD used took place in a virtual environment created in UNITY3D and including the target to localize, a virtual camera, and a tracked pointing tool. Four HTC VIVE base stations were used to track the participants' head and the pointing tool on which HTC VIVE Trackers 2.0 were attached. Participants did not carry a headset and therefore could not explore visually the virtual environment. The pointing task can be separated in several steps that are explained in detail below: the virtual target placement, the video acquisition from a virtual camera, the video processing, the visual-to-auditory conversion and the participants' response collection using the pointing tool.

2.2.1. Virtual target

The virtual target that participants had to localize was a 3D propeller shape of 25 cm in diameter composed of 4 bars with a length of 25 cm and a rectangular section of 5 × 5 cm that was self-rotating at a speed of 10° per video frame (see Figure 1A). The use of an angular shaped target that is self-rotating generated a modification of successive video frames without changing the center position of the target. The orientation of the target was managed in order to continuously face the virtual camera while being displayed. Since participants could not see the virtual target, it was only perceivable through the soundscapes.

2.2.2. Video acquisition

The virtual camera position was set at the beginning of each trial using the position of the head tracker attached on the participants' forehead. Images were acquired with a virtual camera with a field of view of 90 × 74° (Horizontal × Vertical) and a frame rate of 60 Hz. The resulting image was using a grayscale encoding (0–255 gray levels) of a depth map (0.2 m = 0, 5.0 m = 255) of the virtual scene although in this experiment we did not manipulate the depth parameter.

2.2.3. Video processing

Video processing principles are similar to those used by Ambard et al. (2015), aiming to convey only new visual information from one frame to another. Video frames are grayscale images with gray levels ranging from 0 to 255. Pixels of the current frame are only conserved if the gray level pixel-by-pixel absolute difference with the previous frame (frame differencing) is larger than a threshold of 10. The processed image is then rescaled to a 160×120 (Horizontal×Vertical) grayscale image where 0-gray-level pixels are called "inactive" (i.e., no new visual information contained) and the others are "active" graphical pixels (i.e., containing new visual information). Active graphical pixels are then converted into spatialized sounds following a visual-to-auditory encoding in order to generate a soundscape (Figure 2), as explained in the following section.

2.2.4. Visual-to-auditory conversion

The visual-to-auditory conversion consists in the transformation of the processed video stream into a synchronized audio stream that acoustically encodes the extracted graphical features. Each graphical pixel is associated with an "auditory pixel" which is a stereophonic sound with auditory cues specific to the position of the graphical pixel it is associated with. The conversion from a graphical pixel to an auditory pixel follows an encoding that is explained step-by-step in the following sections. Each graphical pixel of a video frame is first associated with a corresponding monophonic audio pixel (detailed in Section 2.2.4.1). The spatialization of the sound using HRTFs is then used to generate a stereophonic audio pixel that simulates a sound source with azimuth and elevation corresponding to the position of the graphical pixel in the virtual camera's field of view (detailed in Section 2.2.4.2). All the stereophonic pixels of a video frame are then compiled to obtain an audio frame (detailed in Section 2.2.4.3). Successive audio frames are then mixed together to generate a continuous audio stream (i.e., the soundscape). Two examples of stereophonic auditory pixels are provided in Figure 2 for each of the three encodings, as well as two examples of soundscapes depending on the location of an object in the field of view of the virtual camera.

10.3389/fpsyg.2023.1079998

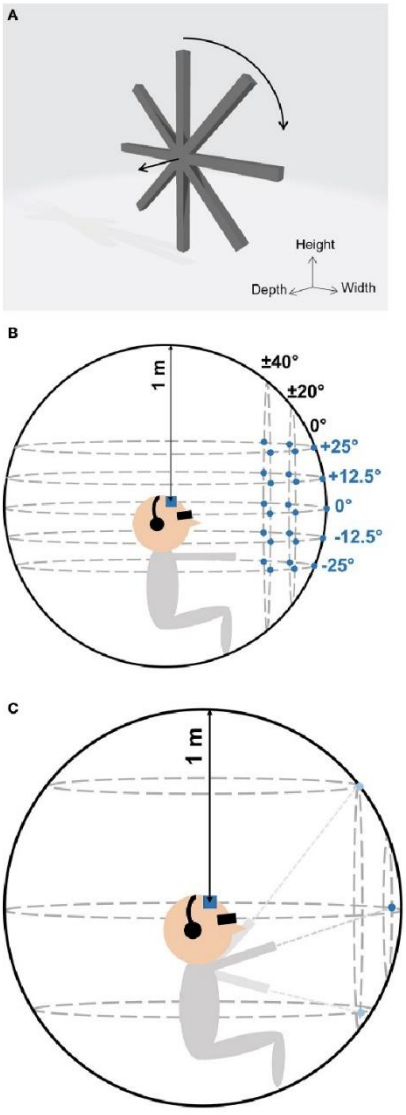


FIGURE 1
(A) The virtual target was a self-rotating 3D propeller shape. The straight arrow shows the forward axis facing the virtual camera. The circular arrow shows the self-rotation direction. **(B)** In each localization test, 25 target positions (blue circles) were tested, including 5 elevation positions (horizontal dotted ellipses) tested at 5 azimuth positions (vertical dotted ellipses). **(C)** In the familiarization session, participants placed the virtual target during 60 seconds on the grid by moving the pointing tool. Participants' head were tracked during the localization tests and familiarization sessions (blue square).

2.2.4.1. Monophonic pixel synthesizing

Three visual-to-auditory encodings were tested in this study: the Noise encoding and 2 Pitch encodings (the Monotonic encoding and the Harmonic encoding). These methods varied in the elevation encoding scheme and in the spectral complexity of the monophonic auditory pixels but all three methods used afterwards the same method for the sound spatialization.

For the Noise encoding, the simulated sound source (i.e., monophonic auditory pixel) in the VAS was a white noise signal generated by inverting a Fourier representation of the auditory pixel with a flat spectrum and random phases.

For the Monotonic encoding, each monophonic auditory pixel was a sinusoidal waveform audio signal (i.e., a pure tone) with a random phase and a frequency related to the elevation of the corresponding graphical pixel in the processed image. For this purpose, we used a linear Mel scale ranging from 344 mel (bottom) to 1,286 mel (top) corresponding to frequencies from 250 to 1,492 Hz.

For the Harmonic encoding, we used the same monophonic auditory pixels as in the Monotonic encoding but instead of a pure tone, we added to it two other frequencies at the 2 following octaves with the same intensity and random phases.

Since the loudness depends on the frequency components of the audio signal, we minimized the differences in loudness between auditory pixels using the *pyloudnorm* Python-package (Steinmetz and Reiss, 2021). Auditory pixel spectrums were then adjusted to compensate for the frequency response of the headphones we used in this experiment (SONY MDR-7506).

2.2.4.2. Auditory pixel spatialization

The azimuth and elevation associated with each pixel were computed based on the coordinates of the corresponding graphical pixel in the camera's field of view. Monophonic auditory pixels were then spatialized by convolving them with the corresponding KEMAR HRTFs from the CIPIC database (Algazi et al., 2001a). This database provides HRTFs recordings with a sound source located in various azimuths and elevations ranging in steps of 5 and 5.625°, respectively. For each pixel, the applied HRTFs were estimated from the database by computing a 4 points time-domain interpolation in which the Interaural Level Difference (ILD) and the convolution signals were separately interpolated using bilinear interpolations before being reassembled as in Sodnik et al. (2005) but using a 2D interpolation instead of a 1D interpolation.

2.2.4.3. Audio frame mixing

Each auditory pixel lasted 34.83 ms including a 5 ms cosine fade-in and a 5 ms cosine fade-out. All the auditory pixels corresponding to the active graphical pixels of the processed current video frame were compiled to form an audio frame. After their compilation, these fade-in and fade-out were still present at the beginning and at the end of the audio frame and they were used to overlap successive audio frames while limiting the artifacts of the auditory transition.

2.2.5. Pointing tool and response collection

The pointing tool was a tracked gun pistol. Participants were instructed to indicate the perceived target position by pointing to it with the gun, with stretched arm. Participants logged their response by pressing a button with their index finger. They were instructed to hold the pointing tool with their dominant hand. The response position was defined as the intersection point of a virtual

10.3389/fpsyg.2023.1079998

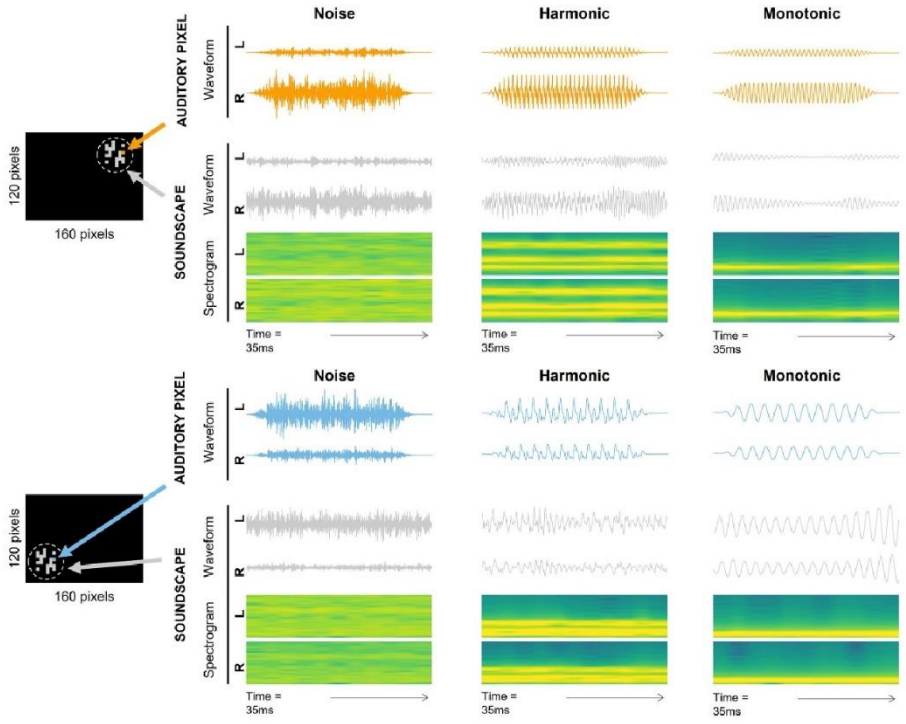


FIGURE 2
Two examples of processed video frames and their corresponding soundscapes. The two processed video frames are depicted in the left side of the figure with a target located on the upper right **(top image)** and bottom left **(bottom image)**. Active and inactive graphical pixels are depicted in gray and black, respectively. Two graphical pixels are highlighted in the video frames (orange in the **top image**, blue in the **bottom image**) and the corresponding auditory pixel waveforms are depicted in the right part of the figure in orange and blue. The corresponding soundscape waveforms (in gray) and soundscape spectrograms of the video frames are also depicted in the right part of the figure. Auditory pixel waveforms, soundscapes waveforms and spectrograms are displayed separately for the Noise encoding **(left column)**, the Harmonic encoding **(middle column)** and the Monotonic encoding **(right column)** and for left (L) and right (R) ear channels separately.

ray originating at the tip of the pointing tool and a virtual 1-m radius sphere with the origin at the location of the virtual camera. The response positions were declined in the elevation response and the azimuth response. The elevation and azimuth signed errors were also computed as the difference between the target position and the response position (in elevation and azimuth separately). A negative elevation signed error indicated a downward shift, and a negative azimuth signed error indicated a shift to the left in the response position. Unsigned errors were computed as the absolute value of the signed errors of each trial.

2.3. Experimental procedure

The experiment consisted in a 45-min session during which participants were seated comfortably in a chair at the center of a room surrounded by the virtual reality tracking system. The participants were equipped with SONY MDR-7506 headphones used to deliver soundscapes. Figure 3 illustrates the timeline of the experimental session. Each participant had to test two visual-to-auditory encodings: the Noise encoding, and a Pitch encoding (Monotonic or Harmonic encoding depending on the group they belonged). Participants from the Monotonic group had to test the Noise encoding and the Monotonic encoding, and participants from the Harmonic group had to test the Noise encoding and the Harmonic encoding. The order of the two tested encodings was counterbalanced between participants so half participants of each group started with the Noise encoding and the other half started with the Pitch encoding.

For each encoding, the participants practiced 2 times the localization test: one without any familiarization or explanation of the encoding and one after a familiarization session. At the beginning of the experiment, participants were instructed to localize a virtual target by pointing to it while being blindfolded. The experimenter

10.3389/fpsyg.2023.1079998

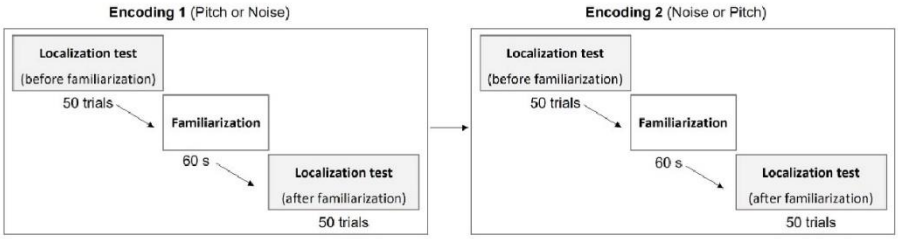


FIGURE 3
Experimental timeline. Participants had to sequentially test the Noise encoding and one of the two Pitch encodings (Monotonic or Harmonic). Participants of the Monotonic and Harmonic groups tested the Monotonic encoding and the Harmonic encoding respectively. For each encoding, participants practiced the localization test two times, before and after a familiarization session.

explained that they will not be able to see the virtual target, but that they will only hear it and that the sound will depend on the position of the target. No indication was given about the way visual-to-auditory encodings worked. Participants were seated and blindfolded using an opaque blindfold fixed with a rubber band and could remove it during breaks. Participants were instructed to keep their head as still as possible during the localization tests. For control purposes, participants' head position was recorded with the tracker every 200 ms to check that they kept their head still. We measured the maximum distance of the head from its mean position for each trial and we found an average maximum distance of approximately 1.5 cm showing that the instructions were rigorously followed.

2.3.1. Localization test

The localization test consisted in 50 trials during which blindfolded participants had to localize the virtual target using soundscapes provided by the visual-to-auditory SSD. During each localization test, the target was located at 25 different positions distributed on a grid of 5 azimuths (−40, −20, 0, +20, and +40°) and 5 elevations (−25, −12.5, 0, +12.5, and +25°). Figure 1B illustrates the grid with the 25 tested positions. As an example, the position [0°, 0°] corresponded to the central position, i.e., the virtual target was centered with the participant's head tracker. For the position [−40°, +12.5°], the target was 40° leftward and 12.5° upward from the central position ([0°, 0°]). The order of the tested positions was randomized and each position was tested 2 times per localization test. The target was placed at 1-meter-distance from the participant's head tracker for all positions (on the virtual 1-meter radius sphere used to collect the response positions).

Each trial started with a 500 ms 440 Hz beep sound, indicating the beginning of the trial. After a 500 ms silent period, the virtual target was displayed at one of the 25 tested positions. Participants were instructed to point with the pointing tool to the perceived location of the target with stretched arm. No time limit was imposed for responding but participants were asked to respond as fast and accurately as possible. The virtual target was displayed until participants pressed the trigger of the pointing tool. The response position was recorded (see Section 2.2.5 for response position computing) and the target disappeared. After a 1,000 ms inter-trial break, the next trial began with the 500 ms beep sound. No feedback was provided regarding response accuracy.

2.3.2. Familiarization session

In between the 2 localization tests of each of the 2 tested encodings, participants practiced a familiarization session which consisted in a 60-s period during which participants freely moved the pointing tool in the front field. Figure 1C illustrates the familiarization session. The virtual target was continuously placed (i.e., no need to press the trigger) on a 1-meter radius sphere centered with the camera position, on the axis of the pointing tool. Consequently, when participants moved their arm, the target was continuously placed at the corresponding position on the 1-meter radius sphere and they could hear the soundscape provided by the encoding corresponding to the processed target images within the camera's field of view. The virtual camera position was updated one time at the beginning of the 60-s timer.

2.4. Data analysis

Statistical analysis were performed using R (version 3.6.1) (Team, 2020). Localization performance during localization tests was assessed separately for azimuth and elevation dimensions, with error-based and regression-based metrics, both fitted with Linear mixed models (LMMs) in order to take into account participants as random factor. All trials of all participants were included in the models without averaging the response positions or the unsigned errors by participant. The LMMs were fitted using the *lmerTest* R-package (Kuznetsova et al., 2017). We used an ANOVA with Satterthwaite approximation of degrees-of-freedom to estimate the effects. *Post-hoc* analysis were conducted using the *emmeans* R-package (version 1.7.4) (Lenth, 2022) with Tukey HSD correction.

2.4.1. Error-based metrics with unsigned and signed errors

Localization performance was assessed through unsigned and signed errors. The elevation signed errors and azimuth signed errors were computed as the difference between target position and response

position in each trial. A negative elevation signed error indicated a downward shift, and a negative azimuth signed error indicated a shift to the left in the response. Only descriptive statistics were conducted on the signed errors. The unsigned errors were computed as the absolute value of the signed error for each trial. They were investigated using LMMs including Encoding (Noise or Pitch), Group (Monotonic or Harmonic) and Phase (Before or After the familiarization) as fixed factors. Therefore, the positions of the target were not included as a factor in the LMMs of the unsigned error. Participants were considered as random effect in both models.

2.4.2. Regression-based metrics with response positions

LMMs were also used for the analysis of the response positions. LMMs included Encoding (Noise or Pitch), Group (Monotonic or Harmonic), Phase (Before or After the familiarization), and Target position as fixed effects. The target elevation only, and the target azimuth only, were included in the elevation response LMM, and in the azimuth response LMM, respectively. Participants were considered as random effect in both models. We used the LMMs predictions to approximate the elevation and the azimuth gains and biases. The gains and biases were obtained by computing the trends (slopes) and intercepts of the models expressing the response position as a function of target position. Note that an optimal localization performance would be obtained with a gain value of 1.0 and a bias of 0.0°.

3. Results

3.1. Performance in elevation localization

The elevation unsigned errors are depicted in Figure 4, left, all target positions combined. Table 1 shows the elevation signed and unsigned errors for each Target elevation, Phase, Encoding and Group. The ANOVA on elevation unsigned errors showed a significant interaction effect of Phase × Encoding × Group [$F_{(1,7556)} = 6.23$, $p = 0.0126$, $\eta_p^2 = 0.0008$]. *Post-hoc* analysis were conducted to investigate the interaction.

The elevation response positions are depicted in Figure 5. The ANOVA showed a significant interaction effect of Phase × Target Elevation × Encoding [$F_{(1,7548)} = 38.84$, $p < 0.0001$, $\eta_p^2 = 0.005$]. We conducted *post-hoc* analysis to investigate the elevation gain (the trend of the model) and bias (the intercept of the model) depending on the Phase and the Encoding. Although the interaction effect of Phase × Target Elevation × Encoding × Group was not significant [$F_{(1,7548)} = 0.50$, $p = 0.48$, $\eta_p^2 = 0.00007$], *post-hoc* analysis were also performed for a control purpose in order to check for differences between the Monotonic and Harmonic groups. The elevation response positions are provided separately for each participant in the Supplementary Figures S1, S2.

3.1.1. Elevation localization performance before the familiarization

Before the practice of the familiarization session, and depending on the encoding, the elevation unsigned errors were comprised between 31.54 ± 27.19° and 40.19 ± 37.02°. For the Monotonic group, the elevation unsigned errors were significantly lower with the Monotonic encoding ($M = 31.54$, $SD = 27.19$) than with the Noise encoding ($M = 40.19$, $SD = 37.03$) [$t_{(7556)} = 7.457$, $p < 0.0001$], suggesting a lower accuracy with the Noise encoding. There was no significant difference in the Harmonic group regarding the elevation unsigned error between the Harmonic encoding ($M = 34.14$, $SD = 33.69$) and the Noise encoding ($M = 35.51$, $SD = 33.69$).

The elevation response positions before the familiarization are depicted in the left panels of the Figures 5A, B for the Monotonic group and the Harmonic group, respectively. The elevation gains were significantly different from 0.0 for all encodings: 0.62 [95% CI = [0.5, 0.74], $t_{(7548)} = 10.118$, $p < 0.0001$] with the Harmonic encoding, 0.61 [95% CI = [0.49, 0.73], $t_{(7548)} = 9.94$, $p < 0.0001$] with the Monotonic encoding, and 0.29 [95% CI = [0.17, 0.41], $t_{(7548)} = 4.728$, $p < 0.0001$] and 0.35 [95% CI = [0.23, 0.47], $t_{(7548)} = 5.746$, $p < 0.0001$] with the Noise encoding of the Harmonic and Monotonic groups, respectively. It suggests that participants could discriminate different elevation positions with the three encodings even before the familiarization.

However, elevation gains were significantly lower than the optimal gain 1.0 with all encodings: with the Harmonic encoding [$t_{(7548)} = -6.173$, $p < 0.0001$], with the Monotonic encoding [$t_{(7548)} = -6.351$, $p < 0.0001$], and with the Noise encoding of the Harmonic group [$t_{(7548)} = -11.562$, $p < 0.0001$] and of the Monotonic group [$t_{(7548)} = -10.544$, $p < 0.0001$]. It depicts a situation where although some variations in elevation seemed to be perceived with the three encodings, participants had difficulties to estimate it before the familiarization.

The participants tended to localize the elevation with a higher performance with the Harmonic or Monotonic encoding than with the Noise encoding. Indeed, the participants from the Harmonic group showed a higher elevation gain with the Harmonic encoding than with the Noise encoding with a significant difference of 0.33 [$t_{(7548)} = -3.811$, $p = 0.0008$]. For the Monotonic group, the elevation gain was also significantly higher with the Monotonic encoding than with the Noise encoding with a difference of 0.26 [$t_{(7548)} = -2.97$, $p = 0.016$]. There was no significant difference regarding the elevation gain between the Harmonic and the Monotonic encodings.

The participants tended to underestimate the elevation position of the targets with the three encodings, as indicated by downward bias and negative elevation errors. In the Monotonic group, the elevation bias were −26.02° (95% CI = [−31.9, −20.16]) with the Noise encoding and −19.52° (95% CI = [−25.4, −13.65]) with the Monotonic encoding. In the Harmonic group, the elevation bias with the Noise encoding and with the Harmonic encoding were −16.33° (95% CI = [−22.2, −10.47]) and −5.73° (95% CI = [−11.6, 0.14]), respectively. With the exception of the Harmonic encoding for which there was just a trend [$t_{(44.9)} = 1.97$, $p = 0.055$], all the elevation bias mentioned above were significantly negative [all $|t_{(44.9)}| > 5.61$, all $p < 0.0001$].

To sum up, participants appeared partially able to perceive a variation of the elevation position of the target with the three encodings before the audio-motor familiarization. Interestingly, participants seemed better able to localize the elevation with the Harmonic and Monotonic encodings.

10.3389/fpsyg.2023.1079998

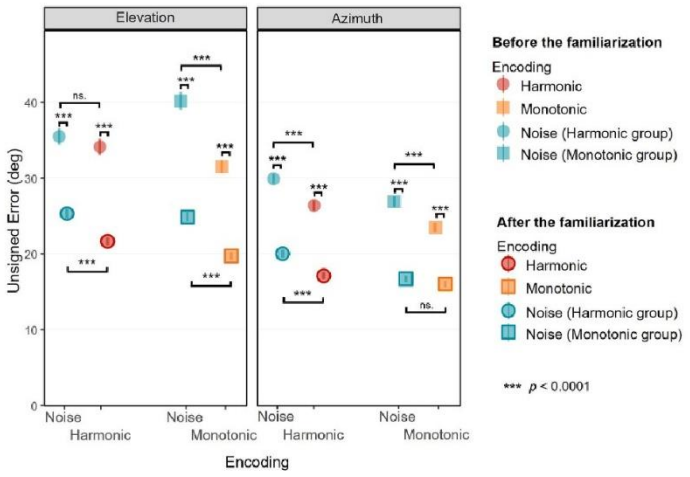


FIGURE 4
Unsigned error in elevation **(left)** and in azimuth **(right)** as a function of the encoding, all target positions combined. Mean unsigned errors (in degree) before (non-surrounded) and after (surrounded) are depicted separately for the Monotonic group (squares) and Harmonic group (circles) and for the three visual-to-auditory encodings: the Noise (blue), the Monotonic (orange) and the Harmonic (red) encodings. Error bars show standard error of the unsigned error.

3.1.2. Elevation localization performance after the familiarization

After the familiarization, the elevation unsigned errors were significantly higher with the Noise encoding than with the 2 pitch-based encodings (Monotonic or Harmonic encodings). With the Noise encoding, the elevation unsigned errors were 24.90 ± 18.40° in the Monotonic group and 25.35 ± 20.31° in the Harmonic group. With the Harmonic and Monotonic encodings, the elevation unsigned errors were 21.70 ± 16.72° and 19.75 ± 16.25° respectively. In the Monotonic group, the elevation unsigned errors were significantly lower with the Monotonic encoding (M = 19.75, SD = 16.25) than with the Noise encoding (M = 24.90, SD = 18.40) [$t_{(7556)} = 4.44$, $p < 0.0001$]. Unlike before the familiarization, the difference was also significant in the Harmonic group. The elevation unsigned errors with the Harmonic encoding (M = 21.70, SD = 16.72) were lower than with the Noise encoding (M = 25.35, SD = 20.31), [$t_{(7556)} = 3.15$, $p = 0.0016$]. Interestingly, the elevation unsigned errors significantly decreased after the familiarization with all the encodings [all $|t_{(7556)}| > 8.75$, all $p < 0.0001$], suggesting that participants localized more accurately the elevation after the familiarization.

The elevation response positions after the familiarization are depicted in the Figures 5A, B for the Monotonic and Harmonic groups, respectively. After the familiarization, the elevation gains were still significantly higher than 0.0 [all $|t_{(7548)}| > 2.9152$, all $p < 0.0036$] with all encodings in the 2 groups. The elevation gains were 1.112 (95% CI = [0.99, 1.23]) with the Harmonic encoding and 1.015 (95% CI = [0.89, 1.14]) with the Monotonic encoding. For the participants of the Harmonic group and the Monotonic group, the elevation gains with the Noise encoding were 0.179 (95% CI = [0.06, 0.30]), and 0.278 (95% CI = [0.16, 0.40]), respectively.

The elevation gains were significantly higher with the Harmonic and Monotonic encodings than with the Noise encoding. We measured a difference of 0.74 [$t_{(7548)} = -8.49$, $p < 0.0001$] in the Harmonic group and a difference of 0.93 [$t_{(7548)} = -10.75$, $p < 0.0001$] in the Monotonic group. Inter-group analysis showed that the difference in elevation gain between the Monotonic and the Harmonic encodings did not significantly differ [$t_{(7548)} = 1.12$, $p = 0.95$].

The elevation gains with the Harmonic and the Monotonic encodings significantly improved after the familiarization to get closer than the optimal gain 1.0. With the Harmonic encoding, the elevation gain significantly increased from 0.62 to 1.112 [$t_{(7548)} = 5.66$, $p < 0.0001$] after which it was not significantly different from the optimal gain 1.0 [$t_{(7548)} = 1.832$, $p = 0.067$]. With the Monotonic encoding, the elevation gain significantly increased from 0.61 to 1.015 [$t_{(7548)} = 4.665$, $p < 0.0001$], and was also no more significantly different from the optimal gain 1.0 [$t_{(7548)} = 0.246$, $p = 0.806$]. However with the Noise encoding in both groups, the familiarization did not improve the elevation gains. In the Harmonic and Monotonic groups, the elevation gains decreased from 0.29 to 0.179 and from 0.35 to 0.278, respectively, but, as previously reported, the decreases were not significant.

Participants kept tending to underestimate the elevation position of the targets with all three encodings, as indicated by persistent negative bias. In the Monotonic group, the elevation bias with the Noise encoding and with the Monotonic encoding were −14.82° (95% CI = [−20.7, −8.96]) and −14.15° (95% CI = [−20.0, −8.28]),

TABLE 1 Elevation signed error and unsigned error (in degree) for each encoding and target elevation, before, and after the familiarization session.

Encoding	Target elevation (degree)	Elevation signed error (degree) Mean ± standard deviation		Elevation unsigned error (degree) Mean ± standard deviation	
		Before familiarization	After familiarization	Before familiarization	After familiarization
Monotonic	+25	−26.71 ± 41.40	−13.21 ± 20.02	37.79 ± 31.55	18.55 ± 15.17
	+12.5	−28.09 ± 33.09	−16.55 ± 22.28	33.45 ± 27.62	21.99 ± 16.89
	0	−18.83 ± 36.41	−11.94 ± 22.73	31.63 ± 26.01	19.60 ± 16.55
	−12.5	−15.14 ± 34.95	−13.23 ± 24.95	27.30 ± 26.50	20.73 ± 19.14
	−25	−8.82 ± 34.34	−15.81 ± 15.18	27.51 ± 22.29	17.89 ± 21.65
Harmonic	+25	−17.66 ± 47.39	−13.73 ± 26.16	36.66 ± 34.76	23.03 ± 18.45
	+12.5	−7.18 ± 45.40	−6.61 ± 26.46	33.41 ± 31.47	20.73 ± 17.67
	0	−5.68 ± 45.71	−10.28 ± 28.07	32.91 ± 32.14	24.06 ± 17.67
	−12.5	−1.10 ± 48.75	−16.43 ± 21.80	33.44 ± 35.40	22.23 ± 15.81
	−25	2.99 ± 48.76	−15.85 ± 16.08	34.28 ± 34.72	18.44 ± 13.01
Noise (Monotonic group)	+25	−42.69 ± 52.91	−33.92 ± 27.00	56.49 ± 37.73	37.54 ± 21.64
	+12.5	−32.93 ± 45.45	−23.24 ± 23.54	43.63 ± 35.23	28.00 ± 17.56
	0	−25.49 ± 43.51	−12.98 ± 24.78	36.61 ± 34.62	23.09 ± 15.73
	−12.5	−20.61 ± 43.35	−7.09 ± 22.23	32.93 ± 34.88	19.02 ± 13.46
	−25	−8.40 ± 47.90	3.11 ± 22.32	31.30 ± 37.15	16.86 ± 14.90
Noise (Harmonic group)	+25	−34.74 ± 46.47	−33.67 ± 28.35	47.86 ± 32.71	36.96 ± 23.87
	+12.5	−24.89 ± 45.12	−21.25 ± 27.08	40.59 ± 31.66	27.87 ± 20.16
	0	−15.68 ± 42.71	−12.20 ± 25.02	32.41 ± 31.87	22.15 ± 16.81
	−12.5	−7.03 ± 44.28	−3.40 ± 25.66	28.84 ± 34.27	19.68 ± 16.76
	−25	0.69 ± 43.84	8.72 ± 25.46	27.84 ± 33.81	20.10 ± 17.85

respectively. In the Harmonic group, the elevation bias with the Noise encoding and with the Harmonic encoding were −12.36° (95% CI = [−18.2, −6.49]) and −12.58° (95% CI = [−18.4, −6.72]), respectively. All the elevation bias were significantly negative [all $|t_{(44.9)}| > 4.24$, all $p < 0.0001$].

To sum up, after the familiarization, the perception of elevation with the Harmonic and Monotonic encodings improved with elevation gains getting closer to the optimal gain. However, the familiarization did not induce any significant improvement in the perception of elevation with the Noise encoding, with persistent low elevation gains in both groups. Additionally, the underestimation elevation bias decreased with the Monotonic and Noise encodings, but not with the Harmonic encoding for which it increased.

3.2. Performance in azimuth localization

The azimuth unsigned errors are depicted in Figure 4, right, all target positions combined. Table 2 shows the azimuth signed and unsigned errors for each Target azimuth, Phase, Encoding and Group. The ANOVA on azimuth unsigned errors showed a significant interaction effect of Phase × Encoding [$F_{(1,7556)} = 5.15$, $p = 0.023$, $\eta_p^2 = 0.00068$], but the interaction including the group was not significant.

The azimuth response positions are depicted in Figure 6. The ANOVA yielded a significant interaction effect of Phase × Target Azimuth × Encoding [$F_{(1,7548)} = 12.69$, $p = 0.0004$, $\eta_p^2 = 0.00005$]. *Post-hoc* analysis were conducted to investigate the azimuth gain (the trend of the model) and bias (the intercept of the model) depending on the Phase and the Encoding. Although the interaction effect of Phase × Target Elevation × Encoding × Group was not significant [$F_{(1,7548)} = 1.64$, $p = 0.20$, $\eta_p^2 = 0.0002$], we conducted *post-hoc* analysis to check for differences between the Monotonic and Harmonic groups for a control purpose. The azimuth response positions are provided separately for each participant in the Supplementary Figures S3, S4.

3.2.1. Azimuth localization performance before the familiarization

Before the practice of the familiarization session, and depending on the encoding, the azimuth unsigned errors were comprised between 23.48 ± 19.48° and 29.91 ± 20.38°. In the Monotonic group, the azimuth unsigned errors were significantly lower with the Monotonic encoding ($M = 23.48$, $SD = 19.48$) than with the Noise encoding ($M = 26.92$, $SD = 22.58$) [$t_{(7556)} = 4.64$, $p < 0.0001$]. The azimuth unsigned errors in the Harmonic group were also significantly lower [$t_{(7556)} = 4.73$, $p < 0.0001$] with the Harmonic encoding ($M = 26.41$, $SD = 20.63$) than with the Noise encoding ($M = 29.91$, $SD = 33.69$).

The azimuth response positions over all participants before the familiarization are depicted in the left panels of the Figures 6A, B for the Monotonic and Harmonic groups, respectively. Before the familiarization, the participants were able to interpret soundscapes to

10.3389/fpsyg.2023.1079998

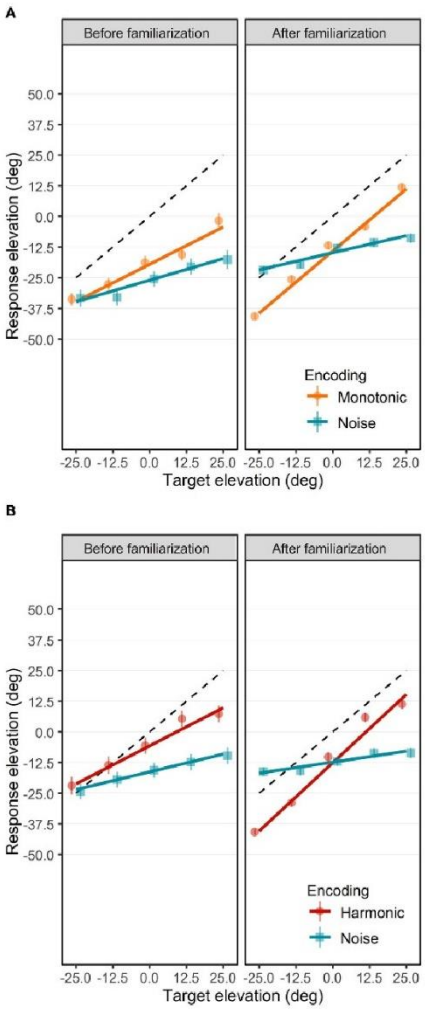


FIGURE 5
Elevation response position as a function of target elevation in the Monotonic group **(A)** and the Harmonic group **(B)**. Mean elevation response positions (in degree) before **(left)** and after **(right)** are represented separately for the three visual-to-auditory encodings: the Noise (blue squares), the Monotonic (orange circles) and the Harmonic (red circles) encodings. Error bars show standard error of elevation response position. Solid lines represent the elevation gains with the Noise (blue), the Monotonic (orange) and Harmonic (red) encodings. Black dashed lines indicate the optimal elevation gain 1.0.

localize the target azimuth. First, the participants perceived different azimuth positions. Indeed, azimuth gains were significantly different from 0.0 with all encodings: 1.81 (95% CI = [1.75, 1.87], $t_{(7548)} = 70.397$, $p < 0.0001$) with the Harmonic encoding, 1.59 [95% CI = [1.54, 1.65], $t_{(7548)} = 61.996$, $p < 0.0001$] with the Monotonic encoding, and 1.88 [95% CI = [1.82, 1.93], $t_{(7548)} = 73.0624$, $p < 0.0001$] and 1.74 [95% CI = [1.68, 1.80], $t_{(7548)} = 67.710$, $p < 0.0001$] with the Noise encoding for the participants in the Harmonic and Monotonic groups, respectively.

Interestingly, the azimuth gains were significantly higher than the optimal gain (i.e., higher than 1.0) with the Harmonic encoding [$t_{(7548)} = 31.492$, $p < 0.0001$], the Monotonic encoding [$t_{(7548)} = 23.091$, $p < 0.0001$], and the Noise encoding in the Harmonic group [$t_{(7548)} = 34.156$, $p < 0.0001$] and the Monotonic group [$t_{(7548)} = 28.804$, $p < 0.0001$]. These gains higher than the optimal gain reflect a lateral overestimation (i.e., left targets localized too much on the left and right targets localized too much on the right) that can be seen with the three encodings.

In the Monotonic group, the overestimation observed with the Noise encoding was significantly higher than with the Monotonic encoding [$t_{(7548)} = 4.04$, $p = 0.0003$]. In the Harmonic group, the overestimation with the Noise encoding compared to the Harmonic encoding was also higher but not significantly. Inter-group comparison of the azimuth gains obtained with the Noise encoding shows a small but significant higher azimuth gain in the Harmonic group [difference of 0.14: $t_{(7548)} = 3.784$, $p = 0.0009$]. As inter-group comparison, we also observed a slight but significant higher overestimation pattern with the Harmonic encoding in comparison with the Monotonic encoding [difference of 0.22: $t_{(7548)} = 5.94$, $p < 0.0001$].

Another interesting result is the tendency to show a left shift as indicated by negative azimuth bias with the three encodings. With the Noise encoding in the Harmonic group, the leftward azimuth bias of −5.03° was significant [$t_{(47.2)} = 2.84$, $p = 0.0066$]. However, leftward azimuth bias with the other encodings were not significantly different from 0.0° (Harmonic group, Noise encoding: −3.23°; Monotonic group, Noise encoding: −2.0°; Monotonic group, Monotonic encoding: −1.233°).

In summary, before the familiarization and with the three encodings, participants were able to localize the azimuth of the target accurately with a tendency to overestimate the lateral eccentricity and a tendency to point too much on the left.

3.2.2. Azimuth localization performance after the familiarization

After the participants practiced the familiarization session, and depending on the encoding, the azimuth unsigned errors were comprised between 16.05 ± 13.38° and 20.05 ± 15.77°. In the Harmonic group, the azimuth unsigned errors were significantly lower with the Harmonic encoding ($M = 17.16$, $SD = 14.97$) than with the Noise encoding ($M = 20.05$, $SD = 15.77$) [$t_{(7556)} = 3.91$, $p = 0.0001$]. The azimuth unsigned errors were not significantly different anymore between the Monotonic encoding ($M = 16.05$, $SD = 13.38$) and the Noise encoding ($M = 16.73$, $SD = 13.50$). Importantly, the azimuth unsigned errors significantly decreased after the familiarization session for all three encodings [all $|t_{(7556)}| > 10.06$, all $p < 0.0001$], suggesting that participants localized more accurately the azimuth after the familiarization.

The azimuth response positions after the familiarization are depicted in the right panels of the Figures 6A, B for the

TABLE 2 Azimuth signed error and unsigned error (in degree) for each encoding and target azimuth, before, and after the familiarization session.

Encoding	Target azimuth (degree)	Azimuth signed error (degree) Mean ± standard deviation		Azimuth unsigned error (degree) Mean ± standard deviation	
		Before familiarization	After familiarization	Before familiarization	After familiarization
Monotonic	+40	17.79 ± 23.73	4.39 ± 18.25	23.16 ± 18.5	13.97 ± 12.5
	+20	25.34 ± 25.50	12.06 ± 18.35	28.71 ± 21.61	17.14 ± 13.69
	0	−7.75 ± 25.39	−5.02 ± 19.33	17.95 ± 19.52	15.21 ± 12.90
	−20	−25.25 ± 21.81	−15.02 ± 18.10	26.93 ± 19.69	18.68 ± 14.27
	−40	−16.27 ± 20.33	−5.34 ± 19.40	20.68 ± 15.79	15.23 ± 13.11
Harmonic	+40	23.26 ± 19.59	5.43 ± 21.49	24.9 ± 17.45	15.83 ± 15.48
	+20	27.17 ± 20.26	12.98 ± 18.80	28.17 ± 19.61	17.23 ± 14.98
	0	−6.89 ± 23.03	−6.02 ± 18.49	16.89 ± 17.36	14.15 ± 12.97
	−20	−32.76 ± 23.38	−16.61 ± 19.93	33.16 ± 22.81	21.6 ± 14.35
	−40	−27.45 ± 23.93	−7.03 ± 22.08	29.25 ± 21.68	16.68 ± 16.05
Noise (Monotonic group)	+40	23.01 ± 25.34	7.27 ± 16.58	27.91 ± 19.79	14.35 ± 11.01
	+20	28.35 ± 29.17	11.96 ± 19.59	30.95 ± 26.38	18.11 ± 14.06
	0	−8.89 ± 24.82	−8.47 ± 14.79	16.78 ± 20.31	12.04 ± 12.05
	−20	−31.27 ± 25.61	−21.37 ± 17.24	33.42 ± 22.71	22.32 ± 15.98
	−40	−21.22 ± 24.10	−11.55 ± 16.88	25.53 ± 19.45	16.82 ± 11.59
Noise (Harmonic group)	+40	24.13 ± 23.85	9.86 ± 24.77	28.65 ± 18.12	19.17 ± 18.49
	+20	29.80 ± 24.77	15.02 ± 17.67	31.78 ± 22.16	18.94 ± 13.35
	0	−11.84 ± 27.02	−7.04 ± 20.76	21.10 ± 20.57	16.34 ± 14.57
	−20	−36.95 ± 20.92	−22.36 ± 18.41	37.28 ± 20.31	25.06 ± 14.48
	−40	−30.29 ± 17.83	−12.40 ± 23.27	30.71 ± 17.09	20.71 ± 16.28

Monotonic and Harmonic groups, respectively. As expected, after the familiarization, participants were still able to localize different azimuth positions by interpreting soundscapes. Azimuth gains were still significantly different from 0.0 with the Harmonic encoding [1.27, 95% CI = [1.22, 1.32], $t_{(7548)} = 49.51$, $p < 0.0001$], with the Monotonic encoding [1.23, 95% CI = [1.18, 1.28], $t_{(7548)} = 47.96$, $p < 0.0001$], and with the Noise encoding for the participants in the Harmonic group [1.41, 95% CI = [1.36, 1.46], $t_{(7548)} = 54.84$, $p < 0.0001$] and in the Monotonic group [1.35, 95% CI = [1.30, 1.41], $t_{(7548)} = 52.711$, $p < 0.0001$], respectively.

The overestimation pattern was still present, as indicated by azimuth gains still significantly higher than the optimal gain 1.0 with all encodings: the Harmonic encoding [$t_{(7548)} = 10.61$, $p < 0.0001$], the Monotonic encoding [$t_{(7548)} = 9.06$, $p < 0.0001$], the Noise encoding in the Harmonic group [$t_{(7548)} = 15.93$, $p < 0.0001$] and in the Monotonic group [$t_{(7548)} = 13.81$, $p < 0.0001$].

Although the lateral overestimation was still significant, it significantly decreased compared to the same localization test before the familiarization. Indeed, the azimuth gains decreased and reached values closer than the optimal gain 1.0 with the 3 encodings. There were significant decreases in azimuth gains of a magnitude of 0.54 [$t_{(7548)} = 14.77$, $p < 0.0001$] and 0.36 [$t_{(7548)} = 9.92$, $p < 0.0001$] with the Harmonic and Monotonic encodings, respectively. The decreases in azimuth gains with the Noise encoding in the Harmonic and the Monotonic groups were also significant with a decrease of a magnitude of, respectively, 0.47 [$t_{(7548)} = 12.897$, $p < 0.0001$] and 0.39 [$t_{(7548)} = 10.61$, $p < 0.0001$].

Additionally, after the familiarization, participants tended to localize the azimuth with a higher performance with the Harmonic and Monotonic encodings than with the Noise encoding. This is suggested by a more pronounced lateral overestimation with the Noise encoding in both groups: the azimuth gains were 0.14 higher [$t_{(7548)} = 3.76$, $p = 0.0042$] and 0.12 higher [$t_{(7548)} = 3.36$, $p = 0.018$] with the Noise encoding in comparison with the Harmonic and Monotonic encodings, respectively.

The slight tendency to show a left shift bias in azimuth was still present with the three encodings. With the Noise encoding in the Monotonic group, the leftward azimuth bias of −4.43° was significant [$t_{(47.2)} = 2.502$, $p = 0.0159$], but in the Harmonic group the bias of −3.38° was just a tendency [$t_{(47.2)} = 1.911$, $p = 0.0621$]. The leftward azimuth bias with the Harmonic encoding (−2.25°) and Monotonic encoding (−1.79°) were also not significant.

To sum up the accuracy in azimuth localization, participants were able to localize target azimuths accurately even before the audio-motor familiarization. After the familiarization, the accuracy increased with a decrease in both the tendency to overestimate the lateral position of lateral targets and the tendency to point too much on the left.

4. Discussion

In this study, we investigated the early stage of use of visual-to-auditory SSDs based on the creation of a VAS (Virtual

10.3389/fpsyg.2023.1079998

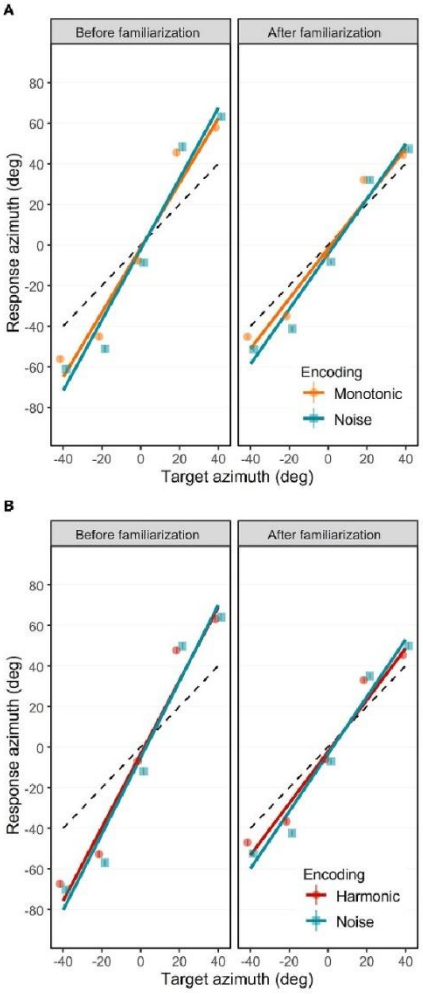


FIGURE 6
Azimuth response position as a function of target azimuth in the Monotonic group **(A)** and the Harmonic group **(B)**. Mean azimuth response positions (in degree) before **(left)** and after **(right)** are represented separately for the three visual-to-auditory encodings: the Noise (blue squares), the Monotonic (orange circles) and the Harmonic (red circles) encodings. Error bars show standard error of azimuth response position. Solid lines represent the azimuth gains with the Noise (blue), the Monotonic (orange) and Harmonic (red) encodings. Black dashed lines indicate the optimal azimuth gain 1.0.

Acoustic Space) for object localization in a virtual environment. Based on soundscapes created using non-individualized HRTFs, we investigated blindfolded participants' abilities to localize a virtual target with three encoding schemes: one conveying elevation with spatialization only (Noise encoding), and two conveying elevation with spatialization and pitch modulation (Monotonic and Harmonic encodings). The two pitch-based encodings varied regarding the sound spectrum complexity: one narrowband with monotones (Monotonic encoding) and one more complex with 2 additional octaves (Harmonic encoding). In order to compare the localization abilities for the azimuth and the elevation with the different visual-to-auditory encodings, we collected the response positions and angular errors of the participants during a task consisting in the localization of a virtual target placed at different azimuths and elevations in their front-field.

4.1. Elevation localization abilities using the visual-to-auditory encodings

4.1.1. Elevation localization performance only based on non-individualized HRTFs is impaired

With the spatialization-based only encoding (Noise encoding), the target was localized before the familiarization with an elevation unsigned error between 27.84 ± 33.81° and 56.49 ± 37.73°. After the familiarization, the elevation unsigned errors decreased to reach values comprised between 16.86 ± 14.90° and 37.54 ± 21.64°. As a comparison, in Mendonça et al. (2013) where the same HRTFs database was used with a white noise sound, the mean elevation unsigned error of participants was 29.3° before practicing a training. The elevation unsigned errors in Geronazzo et al. (2018) without any familiarization and with a white noise sound were comprised between 15.58 ± 12.47° and 33.75 ± 16.17° depending on participants, which is comparable to our results after the familiarization. However, as shown by elevation gains below 0.4 before or after familiarization, the participants had difficulties to discriminate different elevations with this encoding.

The abilities to localize the elevation of an artificially spatialized sound are known to be impaired in comparison with azimuth (Wenzel et al., 1993). Those difficulties arise from the spectral distortions that are specific to individual body morphology (Blauert, 1996; Xu et al., 2007). When using non-individualized HRTFs, these spectral distortions are different from the participant's specific distortions, causing misinterpretation of elevation location. Additionally, the abilities to localize the elevation position of a sound source (virtual or real) are modulated by the spectral content of the sound (Middlebrooks and Green, 1991; Blauert, 1996).

In our study, the difficulty with the spatialization-based only encoding to localize the elevation of the target, even after the audio-motor familiarization, could be explained by a too brief training period to get used to the new auditory cues. Actually, some studies showed an improvement of localization abilities with non-individualized or modified HRTFs after 3 weeks of training in Majdak et al. (2013) or Romigh et al. (2017), or after 2 weeks in Shinn-Cunningham et al. (1998) or 1 week in Kumpik et al. (2010), and about 5 h in Bauer et al. (1966). Moreover, Mendonça et al. (2013) showed the positive long term effect (1-month long) of training in azimuth and elevation localization abilities with a sound source spatialized using the same HRTFs database that was used in the current study. It suggests that the exclusive use of HRTFs to encode spatial information in SSDs might require a long training period or a long process to acquire individualized HRTFs.

4.1.2. Positive effects of cross-modal correspondence on elevation localization

The participants' abilities to localize the elevation of the target using the 2 pitch-based encodings were significantly better than with a broadband sound spatialization encoding. Before the audio-motor familiarization, with the narrowband encoding (Monotonic) and the more complex encoding (Harmonic), the unsigned errors in elevation were comprised between $27.30 \pm 26.50°$ and $37.79 \pm 31.55°$ depending on the target elevation.

Before the familiarization, participants did not receive any information about the way the sound was modulated depending on the target location. In other words, they did not know that low pitch sounds were associated with low elevation locations, and conversely. However, the individual results of each participant for the elevation (Supplementary Figures S1, S2) suggest that even before the familiarization, several participants interpreted the pitch to perceive the target elevation, using high pitch for high elevation and low pitch for low elevation. We suppose that participants were able to guess that the pitch of the sound varied with the target elevation because the experimenter explicitly told them that sound features were modulated as a function of the location of the target although no details regarding this modulation were provided. Two participants (S12 from the Harmonic group and S15 from the Monotonic group) reversed the pitch encoding by associating a low pitch to high elevations and a high pitch to low elevations, but they reversed this miss-representation after the familiarization. Our study showed that after the audio-motor familiarization, the elevation unsigned errors significantly decreased with both pitch-based encodings to reach values comprised between $17.67 \pm 22.23°$ and $24.06 \pm 17.67°$, which are lower elevation unsigned errors than the mean elevation error of 25.2° immediately after the training in Mendonça et al. (2013).

In the visual-to-auditory SSD domain, the artificial pitch mapping of elevation is used by several existing visual-to-auditory SSDs and relies on the audiovisual cross-modal correspondence between visual elevation and pitch (Spence, 2011; Deroy et al., 2018). In the current study, the frequency range was between 250 Hz and about 1,500 Hz with the Monotonic encoding and between 250 Hz and about 6,000 Hz with the Harmonic encoding (i.e., 1,500 Hz × 2 × 2). The floor value of 250 Hz was chosen to provide frequency steps of at least 3 Hz between each of the 120 auditory pixels in a column, to fit to the human frequency discrimination abilities (Howard and Angus, 2009). We used the Mel scale (Stevens et al., 1937) to take into account the perceived scaling in sound frequency discrimination. All the SSDs using a pitch mapping of elevation use different frequency ranges, resolutions (i.e., number of used frequencies) and frequency steps. The vOICe SSD (Meijer, 1992) uses a larger frequency range than the current study (from 500 to 5,000 Hz) following an exponential scale with a 64-frequency resolution. The EyeMusic SSD (Abboud et al., 2014) uses a pentatonic musical scale with 24 frequencies from 65.785 Hz to 1577.065 Hz. The SSD proposed in Ambard et al. (2015) also uses 120 frequency steps but following the Bark scale (Zwicker, 1961) and with a larger frequency range (from 250 Hz to about 2,500 Hz). Technically, increasing the range of frequencies might increase discrimination abilities between target elevations and improve localization abilities. Although, as sound frequency increases the sound feels unpleasant (Kumar et al., 2008). We can postulate that SSD users should be able to modulate some of the parameters in order to adapt the encoding scheme to their own auditory abilities and perceptual preferences.

Our results suggest that a pitch mapping of elevation can quickly be interpreted, even without any explicit explanation of the mapping rules. They also suggest that the pitch mapping provides acoustic cues that are easily interpretable at the early stage of use of a SSD to localize an object. In terms of spatial perception, our study shows that adding abstract acoustic cues to convey spatial information can be more efficient than an imperfect synthesizing of natural acoustic cues. It is difficult to assert that the differences in the results between the Noise encoding and the Pitch encodings are entirely due to the cross-modal correspondence between elevation and pitch since modifying the timbre of the sound by reducing its spectral content also modified how the HRTFs spatialize the sound. Therefore, it would be interesting to investigate the localization performance with monotonic or harmonic sounds in which the pitch is constant (i.e., not related to the elevation of the target) and by conducting an experiment where HRTFs convolution is computed to convey azimuth only, with for instance a constant elevation of 0°.

4.1.3. Insights about the pitch-elevation cross-modal correspondence

Although the aim of this study was not to directly investigate the multisensory perceptual process, the results might bring insights about the pitch-elevation cross-modal correspondence. In the SSD research, it has been suggested that the pitch-based elevation mapping is intuitive in an object recognition task (Stiles and Shimojo, 2015). Based on the results of the current study, it also seems intuitive in a localization task. However, it remains to be further investigated with, for instance, a comparison of elevation localization abilities with a similar pitch-based elevation encoding and another encoding where the direction of the pitch mapping is reversed (i.e., low pitch for high elevation and high pitch for low elevation). The current study also raises the question regarding the automaticity of the cross-modal correspondences as disccussed in Spence and Deroy (2013). In the current study, the facilitation effect of the cross-modal correspondence probably relies on voluntary multisensory perceptual processes. The way the instructions were given to the participants intrinsically induced a goal-directed voluntary strategy in order to infer which modifications in the sound could convey information about the location of the object.

These insights about multisensory process should also be investigated in the blind. Since the pitch-elevation cross-modal correspondence has been suggested to be weak in this population (Deroy et al., 2016), and since auditory spatial perception of the elevation can be impaired in this population (Voss, 2016), it remains to investigate whether similar results would be obtained with blind participants. For this reason, the procedure of the current study was designed in a way to be reproducible with blind participants.

4.1.4. No positive effects of harmonics on elevation localization

The elevation-pitch encoding adds a salient auditory cue while reducing the frequency range where the HRTFs spectrum alterations can operate. To study the effect of the spectral complexity we used an encoding with harmonic sounds (monotonic and 2 following octaves) meant to be a trade-off in terms of spectral complexity between the broadband sound of the Noise encoding and the monotones of the Monotonic encoding. Although pure tones were

used in the Monotonic encoding, it is important to keep in mind that soundscapes were not pure tones. Indeed, soundscapes were made of adjacent auditory pixels, resulting in narrowband but multi-frequency soundscapes (see Figure 2).

The results did not show inter-group differences in the localization accuracy between the Monotonic and the Harmonic encodings. It suggests that adding 2 octaves to the original sound (i.e., the Monotonic encoding) did not modulate the ability to perceive the elevation of the target. Using more complex tones with several sub-octave intervals in the Harmonic encoding might sufficiently modify the sound spectrum to obtain a significant difference with the Monotonic encoding. It could also be interesting to investigate the ability to perceive the elevation of the target with an encoding using sounds containing frequencies higher than the current ceiling frequency (6,000 Hz). However, as mentioned in Section 4.1.1, it seems that the benefits that could arise from the application of the HRTFs on a sound with a broader spectrum could only be perceivable after a long training period.

4.2. Azimuth localization using the visual-to-auditory encodings is accurate but overestimated

Depending on the encoding and the target eccentricity, the magnitude of the azimuth unsigned errors was comprised between $16.78 \pm 20.31°$ and $37.29 \pm 20.32°$. As a comparison, Mendonça et al. (2013) spatialized white noise sounds using the same HRTFs database and their participants localized the azimuth with a mean unsigned error of 21.3° before the training practice. In Geronazzo et al. (2018), the azimuth unsigned errors of participants varied between $3.67 \pm 2.97°$ and $35.98 \pm 45.32°$. In the SSD domain, Scalvini et al. (2022) found a mean azimuth error of $6.72 \pm 5.82°$ in a task consisting in localizing a target with the head. In the current study, after the familiarization, the magnitude of azimuth unsigned errors decreased and was comprised between $12.04 \pm 12.05°$ and $25.06 \pm 14.48°$ depending on the azimuth eccentricity which is comparable to the azimuth unsigned errors found in Geronazzo et al. (2018), without training. In Mendonça et al. (2013), immediately after the training, the mean azimuth unsigned errors also decreased and reached a magnitude of 15.3° which is also comparable to the current results.

In the current study, without any familiarization, and with the three visual-to-auditory encodings, participants were able to discriminate the different azimuths as suggested by gains higher than the optimal value of 1.0. After the familiarization, and with the three visual-to-auditory encodings, participants were able to localize the azimuth of the target with average azimuth gains comprised between 1.23 and 1.41 which were higher than the null value and than the optimal gain 1.0. It shows that the sound spatialization method used in the current study based on HRTFs from the CIPIC database (Algazi et al., 2001a) partly reproduced the natural cues used in free-field sound azimuth localization. These results are not surprising since azimuth is mainly conveyed through binaural cues including the Interaural Level Difference (ILD) and the Interaural Time Difference (ITD) that reflect audio signal differences between the two ears. ITD is mainly used when the spectral content of the audio signal does not include frequencies higher than 1,500 Hz and ILD is mainly used for frequencies higher than 3,000 Hz (Blauert, 1996). The used frequencies ranged from 250 Hz to about 1,500 Hz with the Monotonic encoding, which is in a frequency domain where ITDs are mainly used to perceive the azimuth. With the Harmonic encoding, that added 2 octaves, the frequency range was between 250 and 6,000 Hz which already contains the ILD frequency domain. The Noise encoding with the broadband sound allows both cues (ITD and ILD) to be fully used, which can theoretically improve azimuth localization accuracy in comparison with sounds with a lower spectral complexity, as previously shown in Morikawa and Hirahara (2013). However, in the current study, these drastic changes in the spectrum did not strongly affect the participants' abilities, and the response patterns were similar. In other words, whatever the spectral complexity of the sound used in the encoding (white noise, complex tones or pure tones), binaural cues could be perceived and interpreted by the participants. It can be noticed that azimuth accuracy seems slightly higher with the two pitch-based encodings (the Harmonic and Monotonic encodings) in comparison with the spatialization-based only encoding (the Noise encoding). We did not find similar results in the scientific literature. This facilitation effect could result from a decrease in the cognitive load when the elevation is conveyed through the pitch modulation. As mentioned above, the pitch-based encodings seem more intuitive to localize the elevation, therefore it should globally decrease the cognitive load and thus facilitate the processing of the remaining dimension (i.e., the azimuth dimension). This effect does not seem to drastically shape the results and remains to be confirmed by other experiments.

The participants tended to overestimate the lateral position of the lateral targets with the three visual-to-auditory encodings: a shift to the left for targets on the left, and a shift to the right for targets on the right. Some studies also showed an overestimation pattern of lateral sound sources while using non-individualized HRTFs (Wenzel et al., 1993), in a virtual environment while being blindfolded (Ahrens et al., 2019), using ambisonics (Huisman et al., 2021), and even with real sound sources (Oldfield and Parker, 1984; Makous and Middlebrooks, 1990). A possibility to decrease this overestimation might be to rescale the used HRTF positions to fit to the perceived ones. For example one could rescale the azimuth angles of the HRTFs database to compensate for the non-linear shape that was measured as the perceived ones and measure if it could linearize the response profile.

The participants also tended to localize the targets with a leftward bias between -1.2 and $-5.03°$ in average. This systematic error might be due to a wrong auditory localization but also to a misperception of target distance. Geometrical considerations shows that an underestimation of the distance of the sound source would generate a leftward bias as we see in the current results. Since no indication concerning the sound distance was given, the participants could estimate that the sound sources were located closer than one meter. Figure 7 shows the effect of a misperception of target distance on the azimuth localization. However, for the same reason, a distance underestimation would have cause an overestimation of the elevation perception, which we did not measure.

4.3. A fast improvement in object localization performance

4.3.1. A short but active familiarization method

After a first practice followed by a very short familiarization, participants' abilities to localize an object with the visual-to-auditory

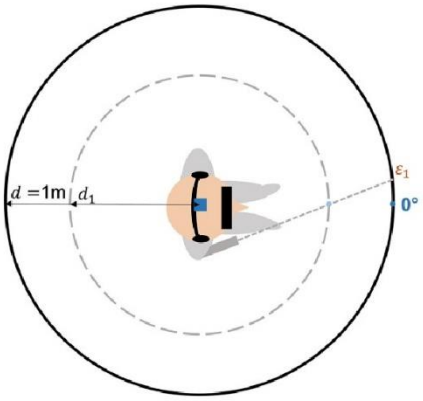


FIGURE 7
Left shift scheme. If the participant perceived the distance of the target closer than the real target distance (d_1 instead of $d = 1$m), it might induce an increase of the leftward bias (ε_1).

SSD were improved. The elevation gains were improved for all the encodings (especially for pitch-based ones), and for the azimuth, the decrease in the lateral overshoot suggests that the interpretation of acoustic cues provided by the ILD and ITD for the azimuth was improved. Since no feedback was given during the first practice, it can be supposed that the familiarization session mainly contributed to acquire sensorimotor contingencies (Auvray, 2004) through the mean of an audio-motor calibration (Aytekin et al., 2008).

In order to avoid a too long experimental session, we used a short audio-motor familiarization session (60 s) during which participants were active by controlling the position of the target, which is known to improve the positive effect of the training (Aytekin et al., 2008; Hüg et al., 2022). Other familiarization methods have been studied and have shown improvements in the use of SSDs. For example, prior to the experimental task, some studies simultaneously displayed to participants an image and its equivalent soundscape (Ambard et al., 2015; Buchs et al., 2021). In another study (Auvray et al., 2007), participants were enrolled in an intensive training of 3 h. Using only a verbal explanation of the visual-to-auditory encoding scheme as been shown to be efficient to understand the main principles of the encoding scheme (Kim and Zatorre, 2008; Buchs et al., 2021; Scalvini et al., 2022). The aim of the current study was not to directly investigate the effect of a short and active familiarization method on localization performance but it shows that a short practice might be sufficient to acquire the sensorimotor contingencies. The effect of the familiarization remains to be clearly assessed by comparing the efficiency of the existing methods with control conditions in order to optimize the SSD learning.

4.3.2. Calibration of the auditory space improves localization abilities

In the current study, participants were not aware of the size of the VAS neither that the head tracker was associated with a virtual camera capturing and converting into sounds a limited portion of the virtual scene in front of them. They only knew that the virtual target would appear at random locations in their front-field at different azimuth and elevation locations. As a consequence, they also did not know the spatial boundaries of the space where the target could be heard. After a short practice, the participants were able to build an accurate mental spatial representation of the virtual space where the visual-to-auditory encoding took place. For instance, the downward bias in elevation decreased after the familiarization session, suggesting that participants learned that the VAS was at a higher location. Also the decrease of the overestimation pattern in azimuth suggests that participants learned that the lateral VAS boundaries were closer.

It has to be noticed that the size of the VAS has an influence on the localization accuracy. The biggest the VAS is, the higher the localization error might be. Restricting the field of view of the camera would result in a smaller possible space in which an heard target could be placed, thus resulting in a lower angular error, but as a counterpart, it would cover a smaller subpart of the front-field without moving the head. For instance, for a target placed in a central position, a random pointing in a VAS with a field of view of 45 × 45° (azimuth × elevation) would result in an error in azimuth and elevation with a standard deviation 2 times lower than with a field of view of 90 × 90° while covering a space 4 times smaller. Studying the effect of various VAS sizes in a target localization task in which the user can freely move the head to point to a target as fast as possible would probably give some insights about the optimal VAS size. However, in ecological contexts, a large VAS size would have the advantage of providing auditory information about obstacles placed with a larger eccentricity with respect to the forward direction of the head. For this reason, in a real context of use, this parameter should probably be customizable according to the habit of use.

5. Conclusion

Long trainings are required to master a visual-to-auditory SSD (Kristjánsson et al., 2016) because the used visual-to-auditory encodings are not enough intuitive (Hamilton-Fletcher et al., 2016b). In our study, we investigated several visual-to-auditory encodings in order to develop a SSD whose auditory information could quickly be interpreted to localize obstacles. In line with previous studies, our results suggest that a visual-to-auditory SSD based on the creation of a VAS is efficient to convey visuo-spatial information about azimuth through soundscapes. Our study shows that a pitch-based elevation mapping can be easily learn to compensate for elevation localization impairments due to the use of non-individualized HRTFs in the creation process of the VAS. Despite a very short period of practice, the participants were able to improve their interpretation of the used acoustic cues both for the azimuth and the elevation encoding schemes.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation. The original contributions presented in the study will be made available in the following link: http://leadserv.u-bourgogne.fr/en/members/maxime-ambard/pages/cross-modal-correspondance-enhances-elevation-localization. Further questions should be directed to the corresponding author.

Ethics statement

The studies involving human participants were reviewed and approved by Comité d'Ethique pour la Recherche de Université Bourgogne Franche-Comté. The patients/participants provided their written informed consent to participate in this study.

Author contributions

CB and MA contributed to conception and design of the study and interpreted the data. CB executed the study and was responsible for data analysis and wrote the first draft of the manuscript in closed collaboration with MA. FS, CM, and JD provided important feedback. All authors have read, approved the manuscript, and contributed substantially to it.

Funding

This research was funded by the Conseil Régional de Bourgogne Franche-Comté (2020_0335), France and the Fond Européen de Développement Régional (FEDER) (BG0027904).

Acknowledgments

Thanks to the Conseil Régional de Bourgogne Franche-Comté, France and the Fond Européen de Développement Régional (FEDER) for their financial support. We thank the Université de Bourgogne and le Centre National de la Recherche Scientifique (CNRS) for the providing administrative support and the infrastructure.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1079998/full#supplementary-material

SUPPLEMENTARY FIGURE 1
Elevation response position as a function of target elevation for each participant of the Monotonic group. Mean elevation response positions (in degree) before **(left)** and after **(right)** are represented separately for the Noise (blue squares) and the Monotonic (orange circles) encodings. Error bars shows standard error of elevation response position. Solid lines represent the elevation gains with the Noise (blue) and Monotonic (orange) encodings. Black dashed lines indicate the optimal elevation gain 1.0.

SUPPLEMENTARY FIGURE 2
Elevation response position as a function of target elevation for each participant of the Harmonic group. Mean elevation response positions (in degree) before **(left)** and after **(right)** are represented separately for the Noise (blue squares) and the Harmonic (red circles) encodings. Error bars shows standard error of elevation response position. Solid lines represent the elevation gains with the Noise (blue) and Harmonic (red) encodings. Black dashed lines indicate the optimal elevation gain 1.0.

SUPPLEMENTARY FIGURE 3
Azimuth response position as a function of target azimuth for each participant of the Monotonic group. Mean azimuth response positions (in degree) before **(left)** and after **(right)** are represented separately for the Noise (blue squares) and the Monotonic (orange circles) encodings. Error bars shows standard error of azimuth response position. Solid lines represent the azimuth gains with the Noise (blue) and Monotonic (orange) encodings. Black dashed lines indicate the optimal azimuth gain 1.0.

SUPPLEMENTARY FIGURE 4
Azimuth response position as a function of target azimuth for each participant of the Harmonic group. Mean azimuth response positions (in degree) before **(left)** and after **(right)** are represented separately for the Noise (blue squares) and the Harmonic (red circles) encodings. Error bars shows standard error of azimuth response position. Solid lines represent the azimuth gains with the Noise (blue) and Harmonic (red) encodings. Black dashed lines indicate the optimal azimuth gain 1.0.

References

Abboud, S., Hanassy, S., Levy-Tzedek, S., Maidenbaum, S., and Amedi, A. (2014). EyeMusic: Introducing a "visual" colorful experience for the blind using auditory sensory substitution. *Restor. Neurol Neurosci.* 32, 247–257. doi: 10.3233/RNN-130338

Ahrens, A., Lund, K. D., Marschall, M., and Dau, T. (2019). Sound source localization with varying amount of visual information in virtual reality. *PLoS ONE* 14, e0214603. doi: 10.1371/journal.pone.0214603

Algazi, V. R, Duda, R., Thompson, D., and Avendano, C. (2001a). "The CIPIC HRTF database," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics* (New Platz, NY:L IEEE), 99–102.

Algazi, V. R., Avendano, C., and Duda, R. O. (2001b). Elevation localization and head-related transfer function analysis at low frequencies. *J. Acoust. Soc. Am.* 109, 1110–1122. doi: 10.1121/1.1349185

Ambard, M., Benezeth, Y., and Pfister, P. (2015). Mobile video-to-audio transducer and motion detection for sensory substitution. *Front. ICT* 2, 20. doi: 10.3389/fict.2015.00020

Asano, F., Suzuki, Y., and Sone, T. (1990). Role of spectral cues in median plane localization. *J. Acoust. Soc. Am.* 88, 159–168. doi: 10.1121/1.399963

Auvray, M. (2004). *Immersion et perception spatiale. L'exemple des dispositifs de substitution sensorielle* (Ph.D. thesis). Ecole des Hautes Etudes en Sciences Sociales, Paris.

Auvray, M., Hanneton, S., and O'Regan, J. K. (2007). Learning to perceive with a visuo–auditory substitution system: localisation and object recognition with 'the voice'. *Perception* 36, 416–430. doi: 10.1068/p5631

Aytekin, M., Moss, C. F., and Simon, J. Z. (2008). A sensorimotor approach to sound localization. *Neural Comput.* 20, 603–635. doi: 10.1162/neco.2007.12-05-094

Bauer, R. W., Matuzsa, J. L., Blackmer, R. F., and Glucksberg, S. (1966). Noise localization after unilateral attenuation. *J. Acoust. Soc. Am.* 40, 441–444. doi: 10.1121/1.1910093

Best, V., Baumgartner, R., Lavandier, M., Majdak, P., and Kopčo, N. (2020). Sound externalization: a review of recent research. *Trends Hear.* 24, 233121652094839. doi: 10.1177/2331216520948390

Blauert, J. (1996). *Spatial Hearing: The Psychophysics of Human Sound Localization, 6th Edn.* Cambridge, MA: MIT Press.

Brown, D., Macpherson, T., and Ward, J. (2011). Seeing with sound? Exploring different characteristics of a visual-to-auditory sensory substitution device. *Perception* 40, 1120–1135. doi: 10.1068/p6952

Buchs, G., Haimler, B., Kerem, M., Maidenbaum, S., Braun, L., and Amedi, A. (2021). A self-training program for sensory substitution devices. *PLoS ONE* 16, e0250281. doi: 10.1371/journal.pone.0250281

Caraiman, S., Morar, A., Owczarek, M., Burlacu, A., Rzeszotarski, D., Botezatu, N., et al. (2017). "Computer vision for the visually impaired: the sound of vision system," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)* (Venice: IEEE), 1480–1489.

Commère, L., Wood, S. U. N., and Rouat, J. (2020). Evaluation of a vision-to-audition substitution system that provides 2D WHERE information and fast user learning. *Techn. Rep. arXiv:2010.09041*, arXiv. arXiv:2010.09041 [cs]. doi: 10.48550/arXiv.2010.09041

Deroy, O., Fasiello, I., Hayward, V., and Auvray, M. (2016). Differentiated audio-tactile correspondences in sighted and blind individuals. *J. Exp. Psychol. Hum. Percept. Perform.* 42, 1204–1214. doi: 10.1037/xhp0000152

Deroy, O., Fernandez-Prieto, I., Navarra, J., and Spence, C. (2018). "Unraveling the paradox of spatial pitch," in *Spatial Biases in Perception and Cognition, 1st Edn*, ed T. L. Hubbard (New York, NY: Cambridge University Press), 77–93.

Evans, K. K., and Treisman, A. (2011). Natural cross-modal mappings between visual and auditory features. *J. Vis.* 10, 6–6. doi: 10.1167/10.1.6

Gardner, M. B. (1973). Some monaural and binaural facets of median plane localization. *J. Acoust. Soc. Am.* 54, 1489–1495. doi: 10.1121/1.1914447

Geronazzo, M., Sikstrom, E., Kleimola, J., Avanzini, F., de Gotzen, A., and Serafin, S. (2018). "The impact of an accurate vertical localization with HRTFs on short explorations of immersive virtual reality scenarios," in *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)* (Munich: IEEE), 90–97.

Hamilton-Fletcher, G., Mengucci, M., and Medeiros, F. (2016a). *Synaestheatre: Sonification of Coloured Objects in Space*. Brighton: International Conference on Live Interfaces.

Hamilton-Fletcher, G., Obrist, M., Watten, P., Mengucci, M., and Ward, J. (2016b). ""I always wanted to see the night sky": blind user preferences for sensory substitution devices," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, CA: ACM), 2162–2174.

Hanneton, S., Auvray, M., and Durette, B. (2010). The Vibe: a versatile vision-to-audition sensory substitution device. *Appl. Bionics Biomech.* 7, 269–276. doi: 10.1155/2010/282341

Hebrank, J., and Wright, D. (1974). Spectral cues used in the localization of sound sources on the median plane. *J. Acoust. Soc. Am.* 56, 1829–1834. doi: 10.1121/1.1903520

Howard, D. M., and Angus, J. (2009). *Acoustics ans Psychoacoustics, 4th Edn*. Oxford: Focal Press.

Hüg, M. X., Bermejo, F., Tommasini, F. C., and Di Paolo, E. A. (2022). Effects of guided exploration on reaching measures of auditory peripersonal space. *Front. Psychol.* 13, 983189. doi: 10.3389/fpsyg.2022.983189

Huisman, T., Ahrens, A., and MacDonald, E. (2021). Ambisonics sound source localization with varying amount of visual information in virtual reality. *Front. Virtual Real.* 2, 722321. doi: 10.3389/frvir.2021.722321

Jicol, C., Lloyd-Esenkaya, T., Proulx, M. J., Lange-Smith, S., Scheller, M., O'Neill, E., et al. (2020). Efficiency of sensory substitution devices alone and in combination with self-motion for spatial navigation in sighted and visually impaired. *Front. Psychol.* 11, 1443. doi: 10.3389/fpsyg.2020.01443

Kim, J.-K., and Zatorre, R. J. (2008). Generalized learning of visual-to-auditory substitution in sighted individuals. *Brain Res.* 1242, 263–275. doi: 10.1016/j.brainres.2008.06.038

Kristjánsson, r., Moldoveanu, A., Jóhannesson, m,. I., Balan, O., Spagnol, S., Valgeirsdóttir, V. V., et al. (2016). Designing sensory-substitution devices: Principles, pitfalls and potential1. *Restor. Neurol Neurosci.* 34, 769–787. doi: 10.3233/RNN-160647

Kumar, S., Forster, H. M., Bailey, P., and Griffiths, T. D. (2008). Mapping unpleasantness of sounds to their auditory representation. *J. Acoust. Soc. Am.* 124, 3810–3817. doi: 10.1121/1.3006380

Kumpik, D. P., Kacelnik, O., and King, A. J. (2010). Adaptive reweighting of auditory localization cues in response to chronic unilateral earplugging in humans. *J. Neurosci.* 30, 4883–4894. doi: 10.1523/JNEUROSCI.5488-09.2010

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). lmertest package: tests in linear mixed effects models. *J. Stat. Softw.* 82, 13. doi: 10.18637/jss.v082.i13

Lenth, R. V. (2022). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.7.4–1.

Levy-Tzedek, S., Hanassy, S., Abboud, S., Maidenbaum, S., and Amedi, A. (2012). Fast, accurate reaching movements with a visual-to-auditory sensory substitution device. *Restor. Neurol Neurosci.* 30, 313–323. doi: 10.3233/RNN-2012-110219

Maidenbaum, S., Abboud, S., and Amedi, A. (2014). Sensory substitution: closing the gap between basic research and widespread practical visual rehabilitation. *Neurosci. Biobehav. Rev.* 41, 3–15. doi: 10.1016/j.neubiorev.2013.11.007

Maidenbaum, S., and Amedi, A. (2019). "Standardizing visual rehabilitation using simple virtual tests," in *2019 International Conference on Virtual Rehabilitation (ICVR)* (Tel Aviv: IEEE), 1–8.

Majdak, P., Walder, T., and Laback, B. (2013). Effect of long-term training on sound localization performance with spectrally warped and band-limited head-related transfer functions. *J. Acoust. Soc. Am.* 134, 2148–2159. doi: 10.1121/1.4816543

Makous, J. C., and Middlebrooks, J. C. (1990). Two-dimensional sound localization by human listeners. *J. Acoust. Soc. Am.* 87, 2188–2200. doi: 10.1121/1.399186

Meijer, P. (1992). An experimental system for auditory image representations. *IEEE Trans. Biomed. Eng.* 39, 112–121. doi: 10.1109/10.121642

Mendonça, C., Campos, G., Dias, P., and Santos, J. A. (2013). Learning auditory space: generalization and long-term effects. *PLoS ONE* 8, e77900. doi: 10.1371/journal.pone.0077900

Mhaish, A., Gholamalizadeh, T., Ince, G., and Duff, D. J. (2016). "Assessment of a visual to spatial-audio sensory substitution system," in *2016 24th Signal Processing and Communication Application Conference (SIU)* (Zonguldak: IEEE), 245–248.

Middlebrooks, J. C. (1999). Individual differences in external-ear transfer functions reduced by scaling in frequency. *J. Acoust. Soc. Am.* 106, 1480–1492. doi: 10.1121/1.427176

Middlebrooks, J. C., and Green, D. M. (1990). Directional dependence of interaural envelope delays. *J. Acoust. Soc. Am.* 87, 2149–2162. doi: 10.1121/1.399183

Middlebrooks, J. C., and Green, D. M. (1991). Sound localization by human listeners. *Annu. Rev. Psychol.* 42, 135–159. doi: 10.1146/annurev.ps.42.020191.001031

Miller, J. (1991). Channel interaction and the redundant-targets effect in bimodal divided attention. *J. Exp. Psychol. Hum. Percept. Perform.* 17, 160–169. doi: 10.1037/0096-1523.17.1.160

Morikawa, D., and Hirahara, T. (2013). Effect of head rotation on horizontal and median sound localization of band-limited noise. *Acoust. Sci. Technol.* 34, 56–58. doi: 10.1250/ast.34.56

Oldfield, S. R., and Parker, S. P. A. (1984). Acuity of sound localisation: a topography of auditory space. I. Normal hearing conditions. *Perception* 13, 581–600. doi: 10.1068/p130581

Pourghaemi, H., Gholamalizadeh, T., Mhaish, A., Duff, D. J., and Ince, G. (2018). Real-time shape-based sensory substitution for object localization and recognition. *Proceedings of the 11th International Conference on Advances in Computer-Human Interactions.*

Proulx, M. J., Stoerig, P., Ludowig, E., and Knoll, I. (2008). Seeing 'where through the ears: effects of learning-by-doing and long-term sensory deprivation on localization based on image-to-sound substitution. *PLoS ONE* 3, e1840. doi: 10.1371/journal.pone.0001840

Real, S., and Araujo, A. (2021). VES: a mixed-reality development platform of navigation systems for blind and visually impaired. *Sensors* 21, 6275. doi: 10.3390/s21186275

Richardson, M., Thar, J., Alvarez, J., Borchers, J., Ward, J., and Hamilton-Fletcher, G. (2019). How much spatial information is lost in the sensory substitution process? Comparing visual, tactile, and auditory approaches. *Perception* 48, 1079–1103. doi: 10.1177/0301006619873194

Romigh, G. D., Simpson, B., and Wang, M. (2017). Specificity of adaptation to non-individualized head-related transfer functions. *J. Acoust. Soc. Am.* 141, 3974–3974. doi: 10.1121/1.4989065

Rouat, J., Lescal, D., and Wood, S. (2014). *Handheld Device for Substitution From Vision to Audition*. New York, NY: 20th International Conference on Auditory Display.

Rusconi, E., Kwan, B., Giordano, B., Umilta, C., and Butterworth, B. (2006). Spatial representation of pitch height: the SMARC effect. *Cognition* 99, 113–129. doi: 10.1016/j.cognition.2005.01.004

Scalvini, F., Bordeau, C., Ambard, M., Migniot, C., and Dubois, J. (2022). "Low-latency human-computer auditory interface based on real-time vision analysis," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Singapore: IEEE), 36–40.

Shinn-Cunningham, B. G., Durlach, N. I., and Held, R. M. (1998). Adapting to supernormal auditory localization cues. I. Bias and resolution. *J. Acoust. Soc. Am.* 103, 3656–3666. doi: 10.1121/1.423088

Sodnik, J., Sušnik, R., Štular, M., and Tomažič, S. (2005). Spatial sound resolution of an interpolated HRIR library. *Appl. Acoust.* 66, 1219–1234. doi: 10.1016/j.apacoust.2005.04.003

Spence, C. (2011). Crossmodal correspondences: a tutorial review. *Attent. Percept. Psychophys.* 73, 971–995. doi: 10.3758/s13414-010-0073-7

Spence, C., and Deroy, O. (2013). How automatic are crossmodal correspondences? *Conscious Cogn.* 22, 245–260. doi: 10.1016/j.concog.2012.12.006

Steinmetz, C. J., and Reiss, J. D. (2021). "Pyloudnorm: a simple yet flexible loudness meter in python," in *150th AES Convention*. Available online at: https://csteinmetz1.github.io/pyloudnorm-eval/

Stevens, S. S., Volkmann, J., and Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *J. Acoust. Soc. Am.* 8, 185–190. doi: 10.1121/1.1915893

Stiles, N. R. B., and Shimojo, S. (2015). Auditory sensory substitution is intuitive and automatic with texture stimuli. *Sci. Rep.* 5, 15628. doi: 10.1038/srep15628

Team, R. C. (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R Core Team.

Voss, P. (2016). Auditory spatial perception without vision. *Front. Psychol.* 07, 01960. doi: 10.3389/fpsyg.2016.01960

Wenzel, E. M., Arruda, M., Kistler, D. J., and Wightman, F. L. (1993). Localization using nonindividualized head-related transfer functions. *J. Acoust. Soc. Am.* 94, 111–123. doi: 10.1121/1.407089

Xu, S., Li, Z., and Salvendy, G. (2007). "Individualization of head-related transfer function for three-dimensional virtual auditory display: a review," in *Proceedings of the 2nd International Conference on Virtual Reality, ICVR'07* (Berlin; Heidelberg: Springer-Verlag), 397–407.

Zwicker, E. (1961). Subdivision of the audible frequency range into critical bands (Frequenzgruppen). *J. Acoust. Soc. Am.* 33, 248–248. doi: 10.1121/1.1908630

Annexe B. Tâche de navigation en environnement virtuel

Résumé

Dans l'objectif d'évaluer les capacités d'utilisation du DSS développé dans la présente thèse mais dans un contexte plus proche d'une utilisation réelle, une étude avec une tâche de navigation sera prochainement conduite. Les données d'un pré-test effectué auprès d'un participant non-voyant sont présentées ci-dessous.

La tâche de navigation prenait place dans un gymnase dans lequel 3 enceintes étaient disposées à 25 m l'une de l'autre en formant un triangle. Le participant non-voyant était équipé d'un casque de réalité virtuelle lui permettant de se déplacer dans un environnement virtuel composé de poteaux virtuel à l'aide du DSS virtuel utilisant le schéma d'encodage de l'Étude 3 de la présente thèse. Dans chacun des 24 essais, la tâche consistait à rejoindre la source sonore réelle émettant un signal sonore le plus rapidement possible en évitant les poteaux. Le champ de vision de la caméra du DSS était de 90°×74° (Horizontal × Vertical) mais deux champs de sonification étaient testés dans deux blocs distincts, se différenciant par le champ de sonification Horizontal et la distance. Le premier champ de sonification était similaire à celui des trois études présentées dans la thèse (horizontal = 90°) mais sur une distance de 2.5 m, et le deuxième était plus étroit (un couloir d'environ 1 m de large mais sur 5 mètres de distance. Chacun des deux blocs était précédé d'une familiarisation audio-motrice avec le schéma d'encodage et le champ de sonification utilisé dans le bloc à suivre. Le nombre de poteaux virtuels par mètre carré dans l'environnement variait de 0 à 0.24/m².

À titre descriptif, les données descriptives préliminaires du participant sont fournies ci-dessous. Le nombre de collision avec un poteau par trajectoire en fonction de la densité de poteau est présenté dans la **Figure B1**. La vitesse de déplacement par trajectoire en fonction de la densité de poteau est présentée dans la **Figure B2**. Les trajectoires parcourues par le participant sont présentées dans la **Figure B3** en fonction de la densité de poteau et du champ de sonification du DSS.

Figure B1. Nombre moyen de collisions avec un poteau virtuel au cours d'un déplacement en fonction de la densité de poteaux virtuels dans l'environnement (peu importe le champ de sonification du DSS). La ligne bleue indique le nombre de collisions qui serait obtenu avec un déplacement aléatoire.

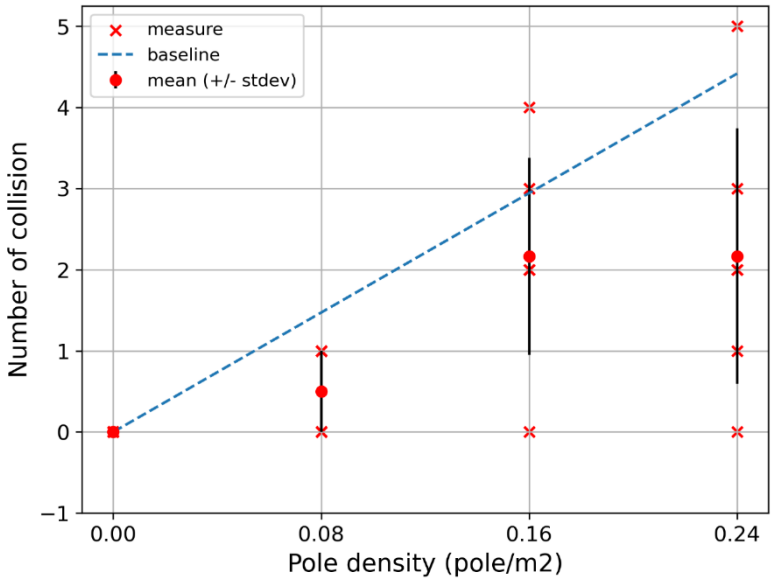


Figure B2. Vitesse moyenne de déplacement (m/s) en fonction de la densité de poteaux virtuels dans l'environnement (peu importe le champ de sonification du DSS).

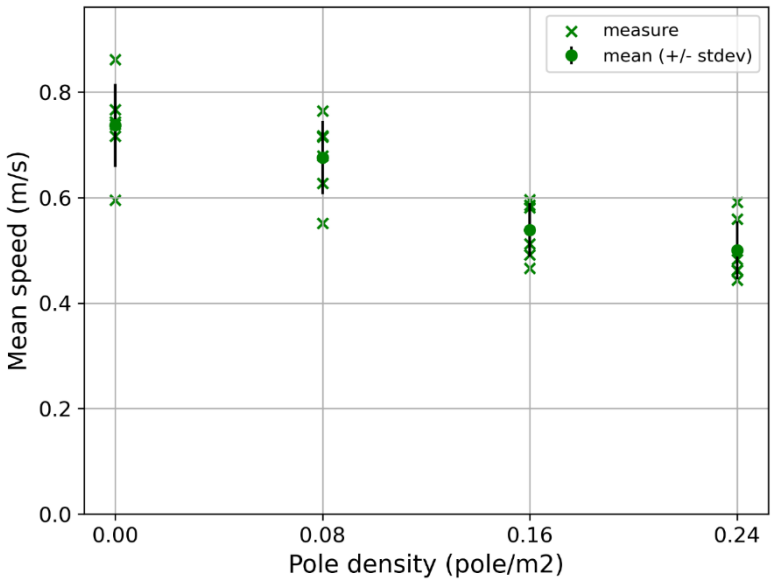


Figure B3. Vue de haut des déplacements du participant non-voyant pendant la tâche de navigation en fonction de la densité de poteaux virtuels dans l'environnement virtuel (*density*= {0/m² ; 0.08/m² ; 0.16/m² ; 0.24/m²} et du champ de sonification du DSS (*FOS* = {FULL : 90° sur 2.5 m ; CENTER = 1 m de large sur 5 m}). Le participant devait rejoindre la source sonore réelle (*Target*, rond bleu), tout en évitant les poteaux (*pole,* points noirs). La trajectoire du participant (*Trajectory*, en vert) et sa vitesse (*speed,* plus la couleur est foncée, plus le participant en lent) ainsi que la zone transmise dans le paysage sonore du DSS durant le déplacement (*Sonified area*, en bleu clair), et les poteaux avec lesquels il est entré en collision (*Collided Pole*, en rouge).

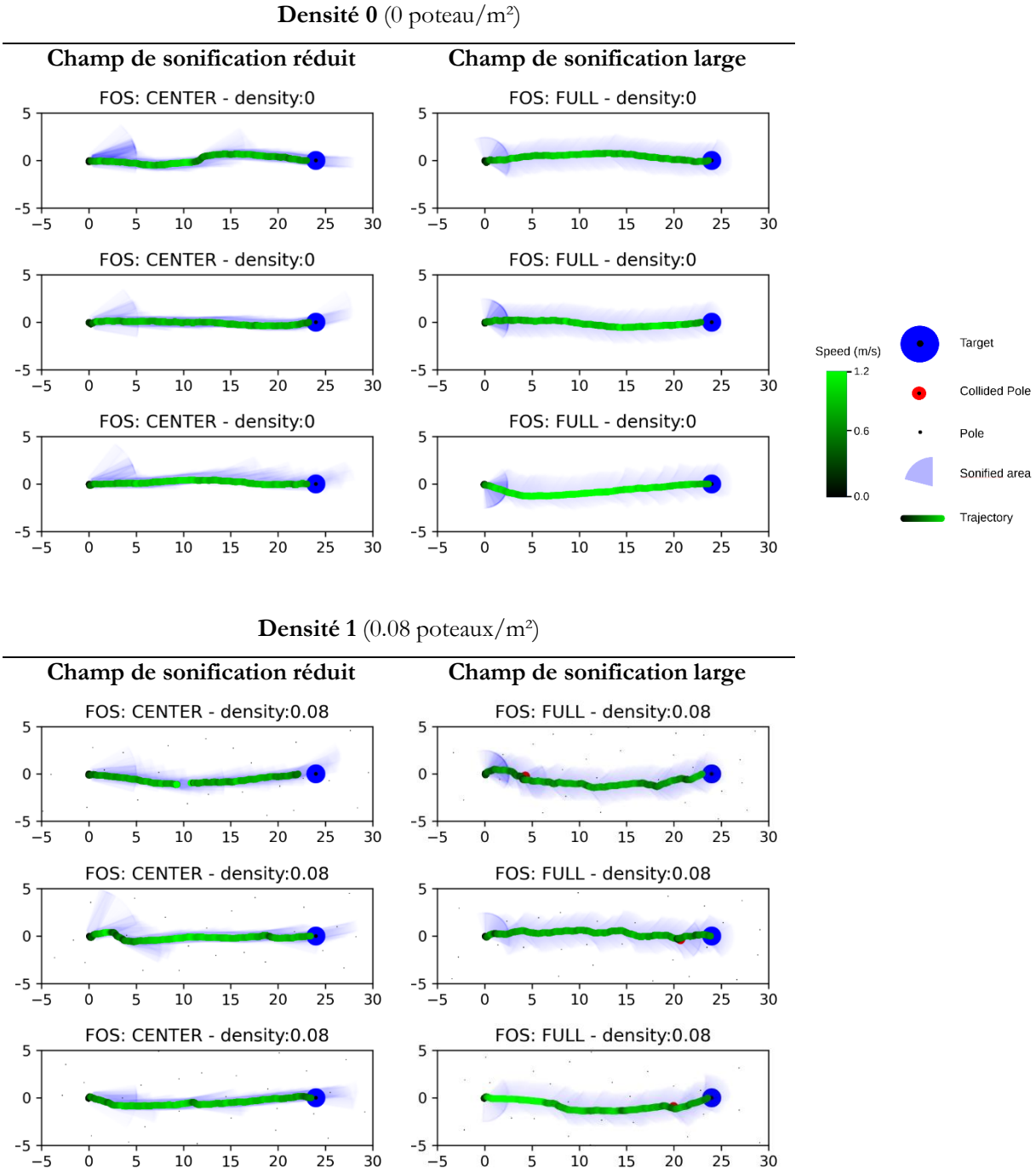
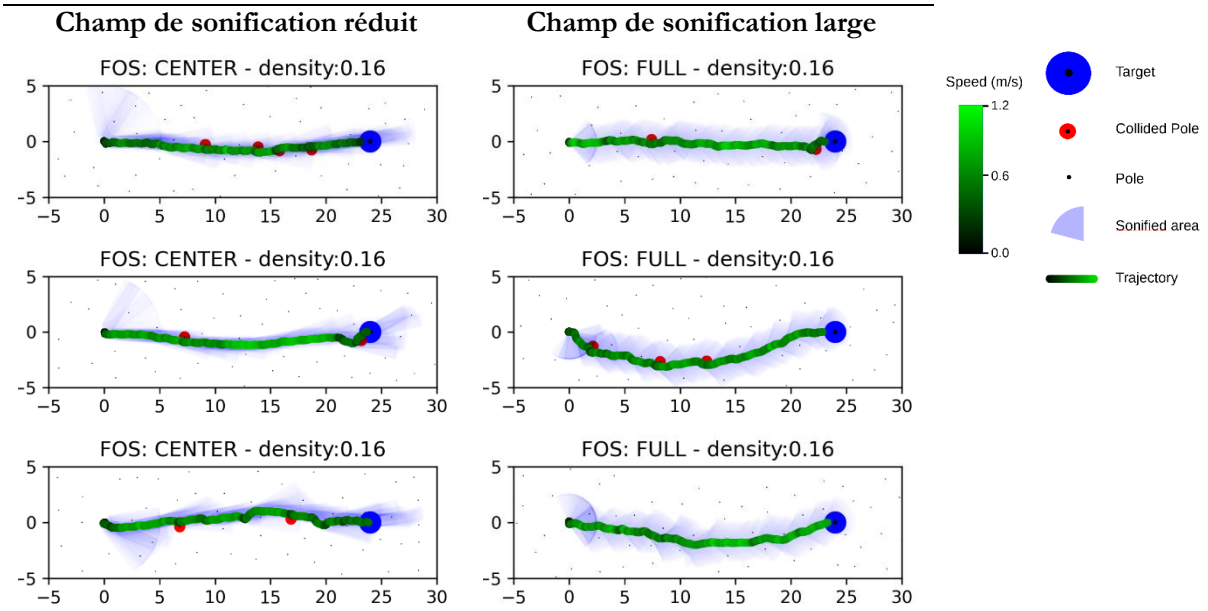
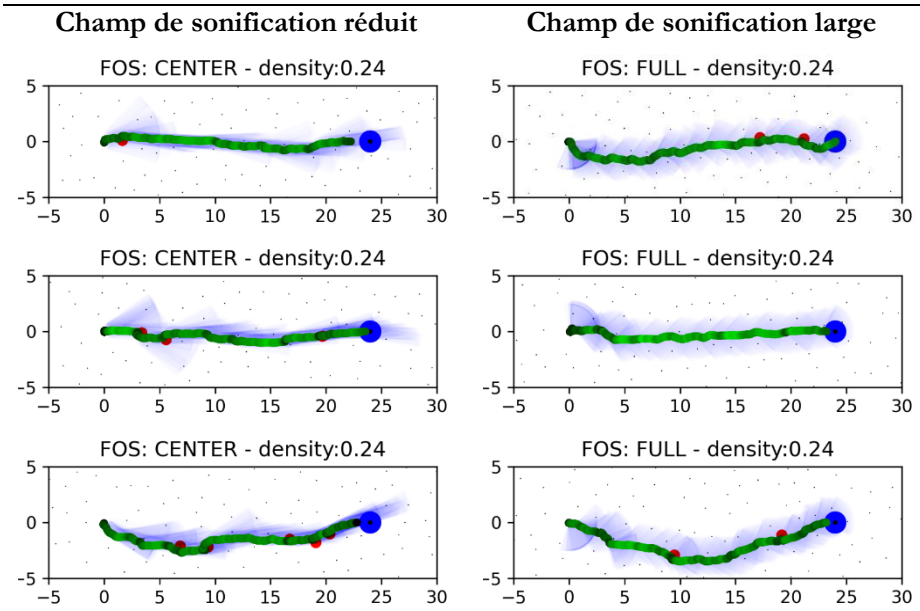


Figure B3. (suite)

Densité 2 (0.16 poteaux/m²)



Densité 3 (0.24 poteaux/m²)



Annexe C. Détection de personnes localisation (ICASSP 2022)

Résumé

Cet article propose une méthode de substitution vision-vers-audition pour aider les personnes non-voyantes à percevoir la scène autour d'elles. Notre approche se concentre sur la localisation des personnes à proximité de l'utilisateur afin de faciliter les déplacements pédestres en milieu urbain. Étant donné qu'une transmission en temps réel et une faible latence sont nécessaires dans ce contexte pour la sécurité de l'utilisateur, nous proposons un système embarqué. Le traitement est basé sur un réseau de neurones à convolutions pour effectuer une localisation 2-dimensionnelle efficace de la personne. Cette mesure est complétée par les informations de profondeur de la personne localisée, puis convertie en un signal acoustique stéréophonique spatialisé avec des fonctions HRTFs. Nous présentons une implémentation basée sur le GPU[3] qui permet un traitement en temps réel à 23 images/s sur un flux vidéo de 640 × 480 pixels sur une ressource matérielle de faible consommation énergétique. Dans une expérience comportementale avec une méthode de pointage avec le corps, nous montrons que cette méthode permet une localisation précise en temps réel d'une personne sur la base des informations auditives du système de substitution.

[3] GPU : Graphics Processing Unit

ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9747094

LOW-LATENCY HUMAN-COMPUTER AUDITORY INTERFACE BASED ON REAL-TIME VISION ANALYSIS

Florian Scalvini[1], *Camille Bordeau*[2], *Maxime Ambard*[2], *Cyrille Migniot*[1], *and Julien Dubois*[1]

[1]ImViA EA 7535 - Univ. Bourgogne Franche-Comté, Dijon, France
[2]LEAD CNRS UMR 5022, Univ. Bourgogne Franche-Comté, Dijon, France

ABSTRACT

This paper proposes a visuo-auditory substitution method to assist visually impaired people in scene understanding. Our approach focuses on person localisation in the user's vicinity in order to ease urban walking. Since a real-time and low-latency is required in this context for user's security, we propose an embedded system. The processing is based on a lightweight convolutional neural network to perform an efficient 2D person localisation. This measurement is enhanced with the corresponding person depth information, and is then transcribed into a stereophonic signal via a head-related transfer function. A GPU-based implementation is presented that enables a real-time processing to be reached at 23 frames/s on a 640x480 video stream. We show with an experiment that this method allows for a real-time accurate audio-based localization.

Index Terms— Auditory sensory substitution, people detection, wearable assistive device, real-time processing

1. INTRODUCTION

A recent study estimates that, despite advances in preventive treatments, the growth and aging of the population should lead to an increase from 43 millions in 2020 to 61 millions of blind people in 2050 [1].

For this people the lack of visual information leads to multiple challenges and daily tasks such as walking in a non-familiar environment without colliding an obstacle remains challenging. The white cane and the trained dog are the classic methods used to remedy this problem. Although these means are very widespread and effective for moving in an unfamiliar environment, they do not provide the user with all the useful information to know his environment. The white cane limits the perception to close objects and the trained dog requires a long training and is relatively expensive.

Assistive technology systems have been the subject of research since decades. Recent advances in image processing methods and the performance of embedded modules in terms of computing power, consumption and miniaturization now offer new possibilities. Among these devices, sensory substitution systems (or SSDs), is a category that uses the brain ability to construct a representation of the world based on a new sensory encoding. These SSDs convert information normally acquired through vision into a signal designed for another sensory modality, mainly auditory or tactile.

For visuo-auditory SSD, this process is called sonification [2]. In the vOICe [3], the pioneer system of visual sensory substitution by sonification, the pixels of a 2D image are sonified according to their positions and luminance.
New sonification protocols provide a stereophonic sound that gives the ability to locate a static or moving object in 2D or 3D scene [4] or to move [5]. The spatial position encoding into a stereophonic sound is computed by Head-Related Transfer Functions simulating the reflections of the sound on the human body before entering the two cochlea. Vision-based and sound generation processing in real-time induce significant latency that should be minimized. For instance setups of some studies [6, 7] acquire data from a smartphone and perform the processing on a laptop in a backpack . Deporting the computing unit to a remote server reduces weight and space requirements and increases autonomy. However, a remote sensing system causes transmission delays and requires a constant connection to operate.

Human-computer auditory interfaces have been designed to localise person [6] based on artificial vision. These systems enable a single class, the obstacles, to be localised in an 2D environment. Alternative methods [8] demonstrate that semantic information could be used in such interfaces as well to perform the environment's perception. The resulting systems enable each significant word to be replaced by discriminant sounds (for instance using different musical instruments). A specific short-sound dictionary has been designed in order to preserve partially the richness of a verbal language meanwhile with respect of the human's physiological reaction-time. Based on this approach, a recent study [2] proposes a SSD using a similar sonification protocol where each object (car, people,...) is associated to a short one-second sound. Nevertheless, a powerful PC platform does not allow real-time to be performed as the algorithm's complexity requires high computational resources. Despite high

Thanks to the Conseil Régional de Bourgogne Franche-Comté, France and the Fond Européen de Développement Régional (FEDER) which are supporting financially this research.

performances in terms of people localisation, the mobility and usability of such a system is then reduced.

In this paper, we propose a human-computer auditory interface system which enables 3D localisation of person based on a RGB-D camera. We focus specifically on person localization since it is one of the most frequently encountered situation during urban walking. Contrary to state-of-art other approaches, the global processing is performed in real-time on standard computational platforms as well as on processing units dedicated to embedded system designs. Hence, a new generation of human-computer auditory interface design is clearly targeted to ease the daily life utilisation and to proposed a mobile and compact system with lower-power consumption, and therefore, higher autonomy. Considering the problems related to a remote transmission, we integrate all video and sound processing on the embedded system.

2. METHOD

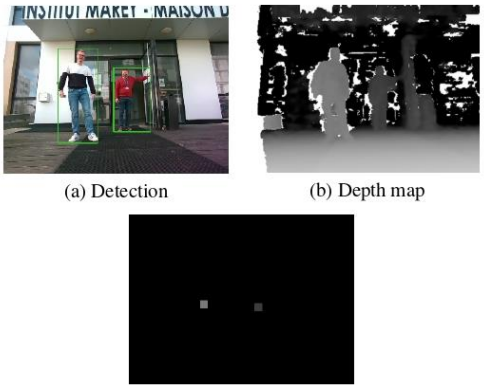
In this section, we describe our method for localizing person in a 3D environment and generating the associated stereophonic sound. Two pipelined processing stages are performed to extract the 3D person localisation.

At the first processing stage: a trained convolutional detection network (CNN) estimates the positions of the persons from the 2D scene. We chose a CNN model based on the architecture You Only Look Once (YOLO) because it has the best framerate [9] for a detection network that produces robust detection on multiple scales. A bounding box detected around the visible part of the person is available as the output of this first processing stage (Figure 1.a). The bounding box size, the 2D positions of the corresponding centroid, and the confidence score are then available for each detection.

The second image processing stage consists in the distance-to-user estimation using the corresponding depth map. This information is provided by the stereoscopic camera (Figure 1.b) which is associated to the rgb standard imaging system.

Finally the position of the centroid is extracted ((Figure 1.c) and sonified with a stereophonic signal generated according to the 3D people localisation. Based on the sound spatialization, the user is then able to localise the targeted person in the 3D space.

Our method of sonification is based on the LibreAudioView system [4] in which each visual object generates an audio signal. In this study, the sonification of the localisation of the target was done as follows: For each video frame the pixel corresponding to the center of the bounding box was extracted if a person has been detected. Given the visual field of the camera, the coordinate of the pixel is mapped on spherical coordinates on a sphere of 2 meter radius centered at the camera. We then used the two meters HRIR data set recorded in an anechoic chamber [10] to spatialize a brief (33ms) monophonic 440 Hz sound with a 5ms cosine fade-in and fade out. For each possible pixel position, we pre-calculated its HRIR spatialization based on the Input responses of the corresponding azimuthal position in the HRIR data set. The amplitude of sound is modulated depending on the distance which separate the target from the user using an inverse square law $A = 1/d^2$ where A is the amplitude and d the distance between the user and the target.



(a) Detection (b) Depth map

(c) Pixels to sonify

Fig. 1. Overview of the method: first a CNN estimates the positions of the person from the color image (a); then the distance-to-user is extracted from the depth map (b) to generate the pixels to sonify (c).

3. REAL-TIME IMPLEMENTATIONS

For our system, we have implemented a CNN trained on the Microsoft Coco database [11]. This database labels 80 different object classes on the 328124 available images. There are 250000 occurrences of person among the 1.5 million of detected objects. Yolov5-small CNN has been selected for implementation in order to propose in fine an embedded system and respect the application constraints. The CNN has been trained on 640×640 resolution images.

The video acquisition is performed using an Intel RealSense D455 stereoscopic camera. Efficient acquisitions can be obtained, equally in outdoor or indoor environments, with a Field of View (FOV) of 87° × 58° and an ideal range of 0.6 to 6m. Both color and depth image were synchronously captured with a resolution of 640×480 pixels at 30 frames per second. The color image is resized to respect the training resolution of the CNN.

As previously mentioned, low-latency system is required in regards to the application constraints. Therefore we have proposed an optimization of the software solution. More

precisely, we aim to minimize the delay between an image acquisition and the sound transmission to the user. The LibreAudioView sonification architecture has been optimized in a previous paper [12]: the optimization of the sonification software has reduced the required processing time by 86% compared to the original version [4]. After the optimization of the sound generation stage, the image processing part represents 95% of the global processing time on a standard PC platform (Intel Core i7-6700HQ processor : 4 Cores - 8 Threads, 2.60 GHz; 16 GB RAM). Therefore, we propose GPU-based implementations to reduce the processing speed. Indeed, the specific multi-core GPU architecture is particularly adapted to regular tasks and to the intrinsic parallelism of the selected algorithm. Finally, we propose a second implementation based on a GPU target that is dedicated to embedded system designs. The goal is to demonstrate that low-latency solution can be designed around such a target to propose a compacted and embedded system. Moreover, the power consumption has been adjusted to increase the system's energy autonomy respecting the application's performance constraints. Different optimisations are then proposed, as the adaptation of the data dynamic, to decrease the system's latency.

First, a comparison is proposed between a standard CPU-based implementation and standard GPU one. As previously, the CPU implementation is based on an Intel Core i7-6700HQ processor (4 Cores - 8 Threads, 2.60 GHz; 16 GB RAM); Meanwhile the GPU implementation is based on a Nvidia GTX 1070 GPU (2048 CUDA Cores, 6.738 Tflops, 8GB VRAM). A laptop is integrating the two targets. The neural network is implemented on both CPU and GPU targets with the Libtorch library (C++ version of Python). The two-first columns of the Table 1 represent the comparison between the two targets and summarizes the performance of the YOLOv5-small and its impact on the overall operation of the sonification device. Please note that the comparison is realized with sequences of images to avoid that the camera frame rate limits the measurement. Considering these results, the inference time of the YOLOv5-small network on a standard computer processor does not enable real-time performances with 640×640 images to be reached whereas on a GPU target this can be achieved. Moreover, the inference time of the CNN on GPU has been reduced by converting the model with the TensorRT SDK. TensorRT is an inference optimizer on Nvidia platforms. A significant gain of 53% between the optimized and non-optimized model is obtained as depicted in the third column of Table 1.

Considering the problems related to a remote transmission, we favour the development of an autonomous device. The low-latency processing is a keystone to reach system's autonomy and hence providing system user's security. A GPU-based implementation represents an appropriate solution to accelerate the processing and therefore a pertinent solution to develop in-fine a wearable device. Indeed, some GPU targets are dedicated to embedded system design by offering a high trade-off between high processing performances and power consumption. Hence, an Nvidia Jetson TX2 (8GB RAM, 256 CUDA Cores, 1.33 TFlops) embedded module has been used. Moreover such target supports and benefits from TensorRT optimization on CNNs. The number of Flops on the embedded card is 5 times lower than on the previous Nvidia GTX 1070 GPU board nevertheless other optimizations are available. Indeed, Nvidia proposes through its Jetson modules and its latest graphics cards (RTX series), the possibility to modify the dynamic range of the CNN's weights. It can be fixed to 16-bits floating format (FP16 : Half precision) instead of 32-bits (FP32). The FP16 configuration decreases significantly the inference time of the CNNs and the memory requirements. On the COCO 2017 validation database, both the configuration provide an accuracy of 55.4% for an overlapping of 50%. Moreover, the selected embedded module, running on Jetpack 4.6 (Ubuntu 18.04, Cuda 10.2, TensorRT 8.0.1), can operate in two power modes: Max-Q of 7.5W (5.5V) and Max-P of 15W. The system's energy autonomy is estimated with a 10 000 mAh - 12 Volts commercial battery. The operating life of the system at full power (TX2 module: 15 W & Realsense camera: 2.335W, the consumption of headphones is ignored) is estimated at 6 hours 55 minutes. The estimated battery life at low power (7.5W + 2.335W) is 12 hours 11 minutes.

	CPU	GPU	
		LibTorch	TensorRT
Yolov5-small (ms)	135	16.3	7.5
Global processing (ms)	142	23.6	15.3

Table 1. Processing time of the YOLOv5-small using a laptop with the overall system (CNN implemented on three targets).

The Table 2 summarizes the impact of the CNN on the embedded target using the two different power modes and considering the two proposed dynamic ranges. An inference of the YOLOv5-small network is lower to the camera framerate on a Nvidia TX2 with a resolution of 640×640 in input.

However the experimental system, with half-precision and maximum power, provides an audio perception equivalent to real-time. The use of this low-power mode generates a larger latency but still enables performances compatible with the application's constrains to be obtained. Hence, this choice of mode is pertinent considering the significant gain in term of operating life.

4. EXPERIMENT

We measured the capabilities offered by such an auditory sensory substitution device in a task consisting in localising human bodies that are in close vicinity. Based on the spatial

	FP32 15W	FP32 7.5W	FP16 15W	FP16 7.5W
Yolov5-small (ms)	55	71	35	47
Global Processing (ms)	62	80	43	56

Table 2. Processing time of the YOLOv5-small on the embedded target in comparison with the overall system. Two power modes and two dynamic ranges are proposed.

information extracted from the 3D video stream, the position of a standing person was transmitted using the spatialized audio encoding described in the section 2.

For this purpose, we used an experimental setup based on an HTC Vive system to track the position of the sensory substitution user's head during a localisation task under two conditions. In the auditory condition, ten blindfolded participants sitting on a 360° revolving chair were placed at the center of a 4m×4m area. A sensory substitution system and a HTC position tracker were fixed on the participant's forehead. The target to localise was a person wearing a second HTC tracker on his sternum, standing at random places 2 meters away from the revolving axis of the chair, on 8 equally spaced positions with a fixed angle gap of 45° as presented in the figure 2.

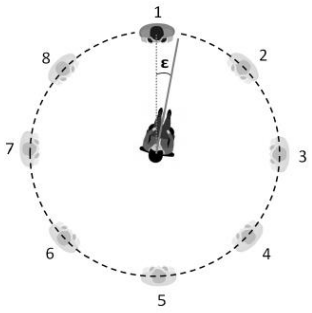


Fig. 2. Experimental setup used to measure the localization abilities. The participant sits at the center of the experiment area on a revolving chair. A standing person randomly changes its standing position among 8 marked places [1...8] equally spaced on a two meters radius circle. We measured the azimuth angle error ε.

Each trial was first composed of 10 seconds of white noise loud enough to cover the sound that might be produced by the target person changing its standing position. After these ten seconds the participant had to rotate on the chair in order to find the target and place it in front of him based solely on the auditory indications provided by the substitution system. The validation was given by pressing on a joystick button. Two trials were performed for each position. In the visual condition, the 10 participants were not blindfolded and the same task was reproduced only with vision, i.e. pointing the head towards the standing person solely using visual feedback.

Results presented in the figure 3 show mean azimuth angular error for each target position in the auditory condition. In this condition, mean azimuth angular error was 6.72° ± 5.82. As expected, mean azimuth angular error in the visual condition was approximately 2 times smaller (2.85° ± 1.99). Despite this difference, these results show that participants could localise a person with high accuracy using our auditory sensory substitution device.

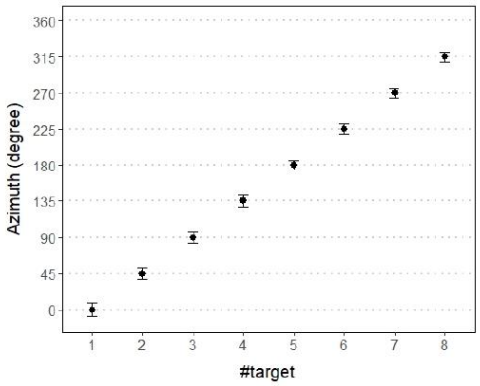


Fig. 3. Azimuth of each target (black dot) with the associated mean value of the azimuth angular error (vertical bar) in the auditory condition.

5. CONCLUSION

In the context of visually impaired people assistance, there are strong constraints in terms of latency, autonomy and portability of the system. In this paper we introduced a new system where the 3D position of person detected by a CNN is sonified into a stereophonic sound. First, tests on laptop have shown that real-time performances with 640×640 images have been achieved on a GPU target. In a wearable device, our system provides an audio perception equivalent or close to real-time. Two power modes and two dynamic ranges allow a compromise between latency and operating life to be adjusted according to the user's preferences. Finally the capacity of a user wearing our device to perceive a person's position has been evaluated and demonstrated experimentally. Future work will extend this protocol to new classes to sonify and enrich the audio signal by a verbal expression of specific events.

6. REFERENCES

[1] Rupert Bourne, "Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis," *The Lancet Global Health*, vol. 5, no. 9, 2017.

[2] Angela Constantinescu, Karin Müller, Monica Haurilet, Vanessa Petrausch, and Rainer Stiefelhagen, "Bring the Environment to Life: A Sonification Module for People with Visual Impairments to Improve Situation Awareness," in *International Conference on Multimodal Interaction*. 2020, pp. 50–59, ACM.

[3] Peter B.L. Meijer, "An experimental system for auditory image representations," *IEEE Transactions on Biomedical Engineering*, vol. 39, no. 2, pp. 112–121, 1992.

[4] Maxime Ambard, Yannick Benezeth, and Philippe Pfister, "Mobile Video-to-Audio Transducer and Motion Detection for Sensory Substitution," *Frontiers in information and communication technologies*, vol. 2, 2015.

[5] Barthélémy Durette, Nicolas Louveton, David Alleysson, and Jeanny Hérault, "Visuo-auditory sensory substitution for mobility assistance: testing TheVIBE," *Workshop on Computer Vision Applicationsfor the Visually Impaired*, pp. 1–13, 2008.

[6] Ruxandra Tapu, Bogdan Mocanu, and Titus Zaharia, "DEEP-SEE: Joint Object Detection, Tracking and Recognition with Application to Visually Impaired Navigational Assistance," *Sensors*, vol. 17, no. 11, pp. 2473 1–24, Oct. 2017.

[7] Matteo Poggi and Stefano Mattoccia, "A wearable mobility aid for the visually impaired based on embedded 3D vision and deep learning," in *IEEE Symposium on Computers and Communication*, Messina, Italy, 2016, pp. 208–213.

[8] Guido Bologna, Benoît Deville, Thierry Pun, and Michel Vinckenbosch, "Transforming 3D Coloured Pixels into Musical Instrument Notes for Vision Substitution Applications," *EURASIP Journal on Image and Video Processing*, vol. 2007, pp. 1–14, 2007.

[9] Andrei-Alexandru Tulbure and Eva-Henrietta Dulf, "A review on modern defect detection models using DCNNs – Deep convolutional neural networks," *Journal of Advanced Research*, 2021.

[10] Hagen Wierstorf, Matthias Geier, and Sascha Spors, "A free database of head related impulse response measurements in the horizontal plane with multiple distances," *Journal of the audio engineering society*, 2011.

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár, "Microsoft COCO: Common Objects in Context," *European Conference on Computer Vision*, pp. 740–755, 2015.

[12] Maxime Ambard, "Software Design for Low-Latency Visuo-Auditory Sensory Substitution on Mobile Devices," *Computer and Information Science*, vol. 10, no. 2, pp. 1, 2017.

Annexe D. Capacités de localisation avec le schéma d'encodage du DSS déterminé au cours de la thèse

Tableau Annexe D. Synthèse des performances de localisation avec le DSS développé dans la présente thèse pour chaque dimension (azimut, élévation et distance) en fonction de la complexité de l'environnement virtuel (Minimaliste : 1 objet ; Complexe : entre 2 et 5 objets) et de la pratique d'un entraînement avant la tâche (protocoles détaillé dans les sections V.1.1 et V.1.2). ***(i)*** schéma d'encodage *Monotonic,* ***(ii)*** schéma d'encodage *INT+ENV*. Les métriques d'erreur de régression sont présentées séparément. La fonction psychophysique est spécifiée pour les métriques de régression, avec p' la position de réponse estimée par le modèle, p la position physique de l'objet, G et B le gain et le biais pour une régression linéaire, et a et k l'exposant et la constante pour une régression puissance. ***(*)*** score de discrimination. ***(**)*** paradigme *open-loop*.

Azimut

Environnement	Entraînement	Étude	Métriques	
			Erreur absolue	Régression $p' = G \times p + B$
Minimaliste	Non	Étude 1 ***(i)***	23.48°	B = -1.23° G = 1.59
Minimaliste	Oui actif	Étude 1 ***(i)***	16.05°	B = -1.79° G = 1.23
		Étude 3	14.2°	B = -3.31° G = 1.81
Complexe	Oui actif	Étude 3	13.0°	B = -3.97° G = 1.65

Elévation

Environnement	Entraînement	Étude	Métriques	
			Erreur absolue	Régression $p' = G \times p + B$
Minimaliste	Non	Étude 1 ***(i)***	31.54 °	B = -19.52° G = 0.61
Minimaliste	Oui actif	Étude 1 ***(i)***	19.75°	B = -14.15° G = 1.02
		Étude 3	15.8°	B = -15.1° G = 0.86
Complexe	Oui actif	Étude 3	15.9°	B = -14.2° G = 0.78

Tableau Annexe D. (suite)

Distance

Environnement	Entraînement	Étude	Métriques	
			Erreur absolue	**Régression** $p' = k \times p^a$
Minimaliste	Oui actif ***(**)***	Étude 2 ***(ii)***	Entre 0.33 et 1.26 m	a = 0.77 k = 0.79
Complexe	Oui actif ***(**)***	Étude 2 ***(ii)***	10.17 cm ***(*)***	.

Annexe E. Base de données (Data in Brief, en cours de révision)

Résumé

La base de données proposée est une collection de séquences vidéos de données de navigation piétonne combinant des informations visuelles et spatiales. Les séquences de navigation piétonne correspondent à des situations rencontrées par un piéton marchant dans un environnement urbain extérieur, comme se déplacer sur le trottoir, se frayer un chemin à travers une foule ou traverser une rue lorsque le feu piéton est vert. Les données acquises sont horodatées et fournissent des images RGB-D associées à un GPS et à des données inertielles (accélération, rotation). Ces enregistrements ont été acquis par des processus distincts, en évitant les retards pendant leur capture afin de garantir une synchronisation entre le moment de l'acquisition par le capteur et le moment de l'enregistrement sur le système. L'acquisition a été réalisée dans la ville de Dijon, en France, dans des rues étroites, des avenues larges et des parcs. Les annotations du capteur RGB-D sont également fournies par des boîtes englobantes indiquant la position d'un objet statique ou dynamique pertinent présent dans une zone piétonne, tel qu'un arbre, un banc ou une personne. Cette base de données de navigation piétonne est proposée pour le développement d'un système d'aide à la mobilité pour les personnes malvoyantes dans leurs déplacements quotidiens en environnement extérieur. Les données visuelles et les séquences de localisation sont utilisées pour élaborer la méthode de traitement de l'image afin d'extraire des informations pertinentes sur l'obstacle et la position actuelle du chemin. Parallèlement à la base de données, une méthode de substitution vision-vers-audition a été utilisée pour convertir chaque séquence d'images en un fichier sonore stéréophonique correspondant. Cet exemple de schéma d'encodage pour la substitution sensorielle vision-vers-audition est fourni dans le but de pouvoir être comparé avec d'autres schémas d'encodage. Des séquences synthétiques associées au même ensemble d'informations sont également fournies sur la base d'enregistrement d'un déplacement dans un modèle 3D d'un lieu réel à Dijon.

Data in Brief
Open access
Article template

ARTICLE INFORMATION

Article title

uB-VisioGeoloc: An image sequences dataset of pedestrian navigation including geolocalised-inertial information and spatial sound rendering of the urban environment's obstacles.

Authors

Florian SCALVINI (1), Camille BORDEAU (2), Maxime AMBARD (2), Cyrille MIGNIOT (1), Mathilde VERGNAUD (1) and Julien DUBOIS (1)*

Affiliations

1 - ImViA EA 7535 – Université de Bourgogne, Dijon, France
2 - LEAD CNRS UMR 5022, Université de Bourgogne, Dijon, France

Corresponding author's email address and Twitter handle

florian.scalvini@u-bourgogne.fr

Keywords

Pedestrian navigation , Virtual scene, Real scene ,Camera RGB-D, GPS, IMU, Sonification, Artificial vision

Abstract

The dataset proposed is a collection of pedestrian navigation data sequences combining visual and spatial information. The pedestrian navigation sequences are situations encountered by a pedestrian walking in an urban outdoor environment, such as moving on the sidewalk, navigating through a crowd or crossing a street when the pedestrian light traffic is green. The acquired data are time-stamped and provide RGB-D images and associated with GPS, and inertial data (acceleration, rotation). These recordings were acquired by separate processes, avoiding delays during their capture in order to guarantee a synchronization between the moment of acquisition by the sensor and the moment of recording on the system. The acquisition was made in the city of Dijon, France, including narrow streets, wide avenues, and parks.
Annotations of the RGB-D are also provided by bounding boxes indicating the position of relevant static or dynamic object present in a pedestrian area such as a tree, bench, or person. This pedestrian navigation dataset is proposed for the development of a mobility support system for visually impaired people in their daily movements in an outdoor environment. The visual data and localization sequences are used to elaborate the visual processing method to extract relevant information about the obstacle and the current position of the path. Alongside the dataset, a visual to auditory substitution method has been employed to convert each image sequence into corresponding stereophonic sound files, allowing for comparison and evaluation. Synthetics sequences associated with the same set of information are also provided based on the recordings of a displacement within the 3D model of a real place in Dijon.

Data in Brief
Open access
Article template

SPECIFICATIONS TABLE

Subject	*Computer Vision : Computer Science Applications*
Specific subject area	*Pedestrian viewpoint image sequences (real & synthetic) dataset with semantic and spatial metadata for computer vision research.*
Data format	The raw color and depth map images are given in uncompressed format (.png). The camera motion files are given in a text file. The filtered annotation files are given in a xml file. The sonified video are given in mkv format encapsulating a video stream encoded with a lossless x264rgb codec and an audio stream encoded with a lossless flac codec.
Type of data	RGB-D image sequences from real and synthetic environments. Annotation of objects on the images. Recording of inertial and GPS sensors. Example of use of these data with sound samples.
Data collection	The dataset consists of two types of data: synthetic and real. The real data was collected using an on-board system comprising an IMU, GPS and RGB-D sensors. The annotation was carried out by a semi-qualified person. The annotation was carried out using a semi-automatic approach combining deep learning techniques and human correction in a post-processing stage. On the other hand, the synthetic data was obtained by navigating in a virtual urban environment generated by a game engine.
Data source location	· *Institution : Université de Bourgogne* · *City/Town/Region: Dijon, Bourgogne-Franche Comté* · *Country: France* · *Latitude and longitude (and GPS coordinates, if possible) for collected samples/data: (47° 19' 19.369" N 5° 2' 29.328" E)*
Data accessibility	Repository name: uB-VisioGeoloc Data identification number: *doi:10.25666/DATAUBFC-2023-07-13* Direct URL to data: https://cloud.u-bourgogne.fr/index.php/s/oaG8cWXXSjQ87Nx

Data in Brief
Open access
Article template

VALUE OF THE DATA

- The dataset is a collection of time-stamped synthetic and real annotated RGB-D image sequences from a pedestrian perspective.
- The real image capture was performed in the urban environment and associated with GPS and inertial data.
- The data contains examples of a visual-to-auditory encoding scheme for 3D visual scene.
- The provided data can be used to apply vision processing method to locate specific objects in an urban pedestrian area with a first-person point of view.
- The inclusion of inertial and GPS data enhances the video processing by providing information on the camera's geographic position, movement, and orientation.
- The proposed dataset could be used by research team to validate and compare a visual substitution device on the same data.

DATA DESCRIPTION

The purpose of this dataset is to provide a comprehensive collection of pedestrian navigation data sequences that reflect everyday life situations encountered in an urban environment. While most existing datasets primarily focus on the driver's perspective [1,2], this dataset aims to capture the pedestrian's point of view and their understanding of the surrounding environment during navigation. The dataset comprises 16 sequences of images that have been collected separately in an urban environment, representing common situations encountered by individuals in everyday life when moving around outdoors. These image sequences represent activities like crossing a road at a pedestrian crossing, moving through crowds, or walking along a pavement next to a road. While these situations may seem harmless, they pose significant challenges for autonomous robots or visually impaired individuals due to the diverse environments and varying levels of complexity. These scenes were either acquired from real-world urban environments or generated using virtual simulations. Both real and virtual environments play complementary roles in providing a comprehensive representation of the challenges faced during pedestrian navigation. The synthetic data, generated using a game engine, allows for precisely controlled and standardized representations of various obstacles. On the other hand, real-world data collected in actual urban environments offers a more diverse and authentic representation of challenges. Additionally, the dataset includes images captured from different camera elevation angles. Figure 1 shows the varying field of view of the camera and its impact on scene acquisition. Specifically, a lower angular elevation prioritizes nearby elements while limiting the maximum capture range. Table 1 gives a concise description of each image sequence, along with the nature of the data and the position of the elevation camera.

Data in Brief
Open access
Article template

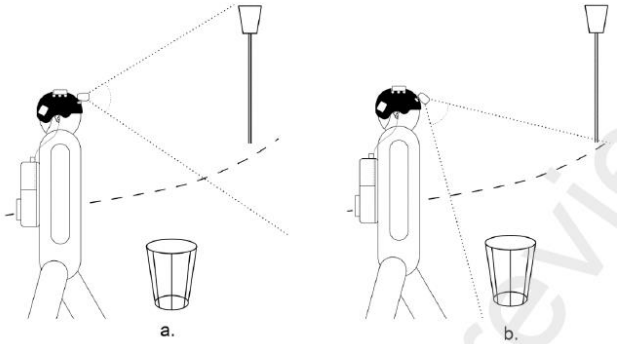


Figure 1: Schematic view illustrating various camera elevation positions. Figure a corresponds to a 0° angle of elevation, while Figure b represents a -40° angle of elevation.

Table 1: Description of each image sequences identified by the index number. The nature column indicates whether the scene is real or virtual. The elevation column provides details about the camera's angular position, with 0° representing a camera parallel to the ground and, the scene description column offers a brief summary of each scene.

Scene	Nature	Elevation	Scene description
1	Synthetic	0°	A person crosses a road to enter a wide pedestrian place occupied by various static obstacles. People is moving around the person while he is continuing to walk on the place.
2	Synthetic	0°	A person is on a wide pedestrian place occupied by various static obstacles and moving persons. He walks 80m within this context while avoiding obstacles.
3	Real	-40°	A person strolling along the designated pedestrian path adjacent to a street, where cars and bicycles are parked. .
4	Real	-40°	A pedestrian walking along the street in the path separated by a line
5	Real	-40°	A person walking on the sidewalk adjacent small street
6	Real	-40°	A person walking on the sidewalk adjacent small street
7	Real	-40°	A person walking on the sidewalk adjacent small street
8	Real	-40°	The pedestrian crosses a main road on a pedestrian crossing with pedestrian light traffic

Data in Brief
Open access
Article template

9	Real	-40°	A bus stops and picks up passengers before leaving
10	Real	-40°	A pedestrian area with tram tracks and a cobblestone path
11	Real	-40°	A pedestrian crossing followed by a large footpath separated from the main road with lots of bikes and people.
12	Real	-40°	The pedestrian walks near small street to join a place near to a high school
13	Real	0°	The pedestrian crosses an important road by multiple crossway
14	Real	0°	The pedestrian walks along a path with bollards, cars and people around him.
15	Real	0°	The pedestrian walks in the city center with temporary signs and pedestrian crossing
16	Real	0°	The pedestrian walks alongside small street to join a place with multiple table and chairs

The image sequences, both real and simulated, provide a combination of visual, semantic, and time-stamped positional information from the pedestrian's perspective. The visual aspect is conveyed by RGB-D images that provide both color and spatial information. Semantic information indicates the location and nature of significant features in the images, while positional data provides information about the viewpoint and displacement of the recording environment. Additionally, audio files are associated with each data subset. These audio files represent the application of a sonification method, designed specifically to aid visually impaired individuals in navigating unfamiliar environments. The audio provides an alternative sensory representation, allowing users to interpret and navigate the environment through sound.

Figure 2: Illustration of image annotations

The dataset consists of RGB images and depth maps recorded at a resolution of 1280 x 720 pixels and a frame rate of 30 frames per second. These specifications were determined by the limitations of the Intel RealSense RGB-D D435i camera used. The depth maps are stored as 16-bit images, representing a distance range of 0 to 65.535 meters. In these depth maps, a difference of value of the brightness

Data in Brief
Open access
Article template

intensity corresponds to 1 millimeter of distance. This information provides detailed spatial understanding of the scene, enabling accurate depth. Each RGB-D image is accompanied by an annotation file that contains semantic information about relevant elements present in the scene. These elements, categorized into 28 classes, include both static and dynamic common objects found in the pedestrian urban environment. The annotation file specifies the locations of these elements within the image by an axis-aligned 2D bounding box (Figure 2).

Dynamic Objects: Car, Bus, Truck, Motorcycle, Scooter, Bicycle, Person, Animal.
Static objects: Tree, Bench, Dining Table, Chair, Fire Hydrant, Garbage, Traffic sign, Traffic light, Pole, Movable sign, Bus station, Traffic light, Pedestrian traffic light, Tree trunk, Crossway, Barricade, Bollard, Bike support, Potted plant.

Table 2: Distribution of the number of images and annotations in each set of recording images

ID	Duration	Frames	Num. of annotation
1	55'	1636	43587
2	54'	1609	30996
3	30'	900	3759
4	30'	900	870
5	30'	900	2558
6	30'	900	8554
7	30'	900	13563
8	53'	1597	13472
9	30'	900	8511
10	26'	804	5449
11	30'	900	36677
12	184'	5522	173539
13	228'	6547	78591
14	285'	8579	40167
15	198'	5899	91848
16	311'	9344	184150

The relevant elements correspond to static or dynamic most common objects present in the pedestrian environment. The position of an object is represented by an axis-aligned 2D bounding box, obtained through a combination of automatic annotation and manual verification. The annotation file format follows the widely accepted Pascal VOC format used in object recognition datasets. Table 2 provides a summary of the characteristics of the 16 image sequences within the dataset. The Figure 3 illustrates the occurrence of each class in the dataset, with color indicating groups of classes sharing similar characteristics and color nuances representing synthetic (Pastel color, top of the column) or real data segments (bright colors, lower part).

Data in Brief
Open access
Article template

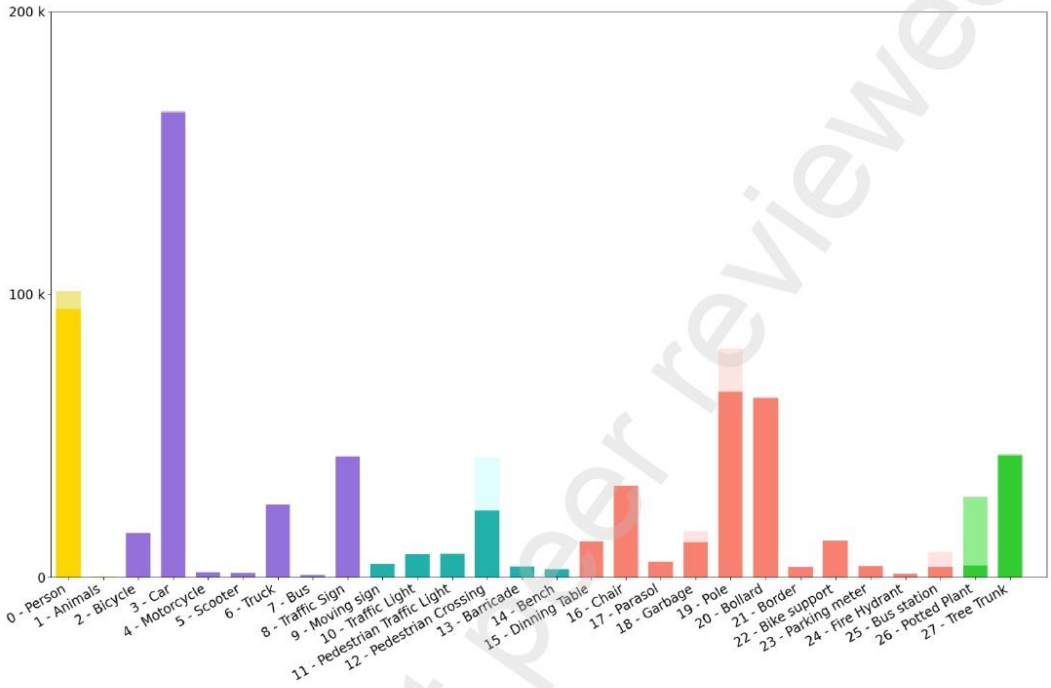


Figure 3: Class distribution in the dataset

The dataset includes recorded position, motion and orientation data from a navigational perspective. This information is captured either directly in the virtual environment or by an IMU (Inertial Measurement Unit) sensor and GPS antenna in the recording system. The recorded data is stored in text files, with each line representing a set of time-stamped data separated by semicolons. Each line includes the recording time (in the format HH:MM:SS:SSS), the orientation of the viewpoint expressed as a quaternion, and the position information. The position is given in GPS coordinates. In case of the real scenes, these values are provided by a GPS sensor (type of sensor) at a rate of 10Hz. In case of the synthetic scenes, these GPS values were computed based on the recorded displacements on the scene projected with an equirectangular method to retrieve GPS coordinates. Additional information is provided in the form of the gravity (m/s^2) and the magnetic (uT) vectors.

EXPERIMENTAL DESIGN, MATERIALS AND METHODS

Despite the similar data formats, the methods of data acquisition and annotation differed depending on whether the data came from virtual space or the real world. The synthetic data was generated using the virtual camera of the Unity game engine and annotated directly, while the real data was captured using an RGB-D camera and annotated using a convolutional neural network based on deep learning. This section explains in detail the techniques used for the acquisition and annotation of the videos, and describes the method used to produce the accompanying sound files.

Data in Brief
Open access
Article template

a. Real data

The data collection process in the city of Dijon involved a person equipped with a tracking system. The system, mounted on the user's helmet at a height of 1.85 meters, replicated the visual perspective of the person. The elevation of the camera was adjusted using the fixation screw, allowing the camera angle to be fine-tuned. The capture system, as shown in Figure 4, consisted of several components. The primary component was the Intel Realsense D435 RGB-D camera, which captured both color (RGB) and depth (D) information at 30 frames per second. This camera was responsible for capturing the visual data as the person navigated. In addition to the camera, the acquisition system included an Adafruit BNO055 Inertial Measurement Unit (IMU) sensor and a GPS antenna, which recorded information about the environment at a frequency of 100 Hz and 10 Hz respectively. An Adafruit MCP2221A UART to USB module is also part of the device to facilitate communication between the IMU and GPS sensors and the laptop. Data acquisition was performed by a C++ program using the Realsense 2 and OpenCV libraries. A multi-threaded approach was used to ensure synchronization between the data acquisition rate and the recording process. Each sensor had its own dedicated thread running independently of the others. This design allowed each sensor to operate autonomously, unaffected by the performance of the other threads.

Figure 4: Experimental acquisition setup

The annotation of the relevant elements of the urban scene according to the list of classes was generated using a semi-automatic method. This method involved a combination of pre-annotation by a convolutional neural network (CNN) and human annotation to ensure accurate labeling, including missing classes and correcting mislabeled elements. The pre-annotation stage utilized a YOLOv7 [5] architecture, which was trained on two distinct datasets. These datasets were carefully selected to cover a wide range of objects commonly found in urban areas. The MS COCO dataset provided annotations for 80 common objects, encompassing outdoor items, animals, and more. On the other hand, the SideGuide dataset specifically focused on pedestrian obstacles in a South Korean city. Additionally, a subset of annotated images from our dataset was used for training. However, to ensure the accuracy of the labels, a human correction and annotation step was applied. This manual process involved checking and adjusting the annotations frame by frame using a standard annotation tool.

Data in Brief
Open access
Article template

b. Synthetic data

The synthetic environment models the popular Darcy place of the city of Dijon, France. This low-poly virtual model was done based on dot clouds acquired using a LiDAR scan (Light Detection And Ranging). The trajectory of the head was recorded by wearing a virtual reality headset (Oculus Quest 2.0) in an empty gymnasium. The 3D model and the head trajectories were reused in the 3DSmax software for optimal graphical rendering. The bounding boxes of the objects were determined by the axis align portion of the screen were the object is visible. To avoid the multiplicity of very small bounding boxes of objects situated far from the camera but still visible on the screen, we only kept the objects that were less than 50m away from the camera and with a bounding box of more than 50 pixels. The same list of class than for the real data was used for the annotation of the relevant elements. The annotation was generated automatically with a C# script running in the Unity software using the labeling of the 3D virtual objects.

c. Sound generation

An application of visual-to-auditory encoding scheme on the dataset was performed with associated audio files. The encoding scheme based on the *Monotonic* encoding from [6] consists in an image processing step followed by an image-to-sound conversion. The video processing extracts the brightness intensity variation of the pixel by differentiating two successive depth maps previously remapped into 8-bit images and rescaled between 0.2m and 5.2m with a resolution of 160 x 120 pixels. The range was selected to focus on nearby elements that could result in a danger for visually impaired people. The resulting image-to-sound conversion is based on the association of the pixel position and brightness with a unique 3D spatialized sound, where the encoding scheme combines information about elevation, azimuth, and distance. The spatialized sound is a pure tone whose frequency depends on the elevation (from 250 Hz to 1492 Hz) that is convolved with HRTFs from the CIPIC database [7] to obtain a spatialized stereophonic sound in azimuth and elevation. Finally, the distance encoding is added by modulating the sound intensity and the envelope amplitude as a function of the pixel brightness intensity. All the generated sounds were combined to produce an audio frame, and then the process is repeated until the end of the set of images to obtain a sequence of audio frames that generates an audio stream.

LIMITATIONS

This data set has certain limitations that need to be taken into account. Data acquisition during pedestrian navigation required an on-board system where the sensors are connected via a USB connection to a laptop and introduced a potential for lost images (< 5%). The presence of missing images is due to the preference for acquiring high-resolution images, which are more advantageous for the research team. These missing images, although corrupted, are included to avoid confusion. Moreover, the IMU accelerometer data included in the dataset exhibits drift over time, making it unsuitable in its raw form for accurately tracking the user's position over an extended period of time, but may nevertheless allow displacement to be estimated over a short period. Some incorrectly labelled images can persist even if the data has been carefully annotated by several people who have examined the same image several times. On the other hand, the synthetic data from the dataset provides very accurate IMU and GPS information, enabling precise mapping on a 2D map of Dijon.

Data in Brief
Open access
Article template

ETHICS STATEMENT

Ethical considerations are of utmost importance when dealing with a dataset of visual information acquired in a real urban environment. Indeed, the collection, the diffusion and the use of such data should follow short ethical guidelines to ensure the protection of privacy. In our dataset, we adhere to French regulations regarding the recording of scenes in public spaces without explicit consent. The French regulations requires that the person shouldn't be identifiable if there are in the foreground of the scene. Consequently, when a person is relatively close of the camera, we changed slowly the field of view of the camera on another element or on the ground. In this way, the person's face was not recorded without their consent.

CREDIT AUTHOR STATEMENT

***Florian SCALVINI**: Conceptualization, Methodology, Software, Writing - Original Draft. **Camille BORDEAU**: Conceptualization, Methodology, Writing - Original Draft. **Maxime AMBARD**: Methodology, Software, Validation, Writing - Review & Editing. **Cyrille MIGNIOT** : Writing - Review & Editing. **Mathilde VERGNAUD**: Resources, Data Curation. **Julien DUBOIS**: Writing - Review & Editing, Supervision*

ACKNOWLEDGEMENTS

Thanks to the Conseil Régional de Bourgogne Franche-Comté, France and the Fond Européen de Développement Régional (FEDER) which are supporting financially this research.

DECLARATION OF COMPETING INTERESTS

- The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

1. Geiger A, Lenz P, Stiller C and Urtasun R. Vision meets robotics: The KITTI dataset. The International Journal of Robotics Research. 2013;32(11):1231-1237. doi:10.1177/0278364913491297
2. M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes Dataset for Semantic Urban Scene Understanding," in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
3. K. Park et al., "SideGuide:A Large-scale Sidewalk Dataset for Guiding Impaired People," 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 2020, pp. 10022-10029, doi: 10.1109/IROS45743.2020.9340734.
4. Lin, TY. et al. (2014). Microsoft COCO: Common Objects in Context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8693. Springer, Cham. https://doi.org/10.1007/978-3-319-10602-1_48
5. Chien-Yao Wang, Alexey Bochkovskiy, Hong-Yuan Mark Liao; "YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7464-7475
6. Bordeau, C., Scalvini, F., Migniot, C., Dubois, J.,Ambard, M. (2023). Cross-modal correspondence enhances elevation localization in visual-to-auditory sensory substitution. Frontiers in Psychology, 14. https://doi.org/10.3389/fpsyg.2023.1079998
7. Algazi, V. R., Duda, R. O., Thompson, D. M., Avendano, C. (2001). The CIPIC HRTF database. Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575), 99-102. https://doi.org/10.1109/ASPAA.2001.969552

Annexe F. Guidage et détection d'obstacle (SITIS 2022)

Résumé

Cet article propose un nouveau dispositif d'assistance à la mobilité pour aider les personnes malvoyantes à atteindre une destination définie en toute sécurité dans un environnement intérieur. Cette approche, basée sur la substitution sensorielle vision-vers-audition, transmet à l'utilisateur, à la fois la destination à atteindre, et les obstacles proches et dangereux, par le biais d'informations auditives spatiales 2-dimensionnelles. Des marqueurs visuels sont placés à plusieurs endroits pertinents du bâtiment pour créer un maillage du bâtiment où chaque marqueur est visuellement accessible à partir d'un autre marqueur. Un algorithme de recherche d'itinéraire permet de définir le chemin le plus court pour atteindre la position souhaitée. La tâche de navigation est réalisée en se déplaçant de marqueur visuel en marqueur visuel jusqu'à ce que la destination souhaitée soit atteinte. Ces marqueurs peuvent être utilisés indépendamment de tout autre système, ou en complément d'autres solutions basées sur la géolocalisation et/ou sur un modèle numérique du bâtiment. De plus, d'autres informations peuvent être associées aux marqueurs et verbalisées à l'utilisateur, comme un danger temporaire, la présence d'une porte ou toute autre information habituelle. Ces marqueurs visuels passifs permettent de déployer facilement, et rapidement, une solution évolutive et peu coûteuse pour signaler l'environnement aux utilisateurs. Combinée à notre détection d'obstacles en temps réel, leur analyse permet d'améliorer les capacités de navigation des personnes malvoyantes.

Visual-auditory substitution device for indoor navigation based on fast visual marker detection

Florian SCALVINI
ImViA EA 7535
Univ. Bourgogne-Franche-Comté
Dijon, France
florian.scalvini@u-bourgogne.fr

Camille BORDEAU
LEAD CNRS UMR 5022
Univ. Bourgogne-Franche-Comté
Dijon, France
camille.bordeau@u-bourgogne.fr

Maxime AMBARD
LEAD CNRS UMR 5022
Univ. Bourgogne-Franche-Comté
Dijon, France
maxime.ambard@u-bourgogne.fr

Cyrille MIGNIOT
ImViA EA 7535
Univ. Bourgogne-Franche-Comté
Dijon, France
cyrille.migniot@u-bourgogne.fr

Stéphane ARGON
LEAD CNRS UMR 5022
Univ. Bourgogne-Franche-Comté
Dijon, France
stephane.argon@gmail.com

Julien DUBOIS
ImViA EA 7535
Univ. Bourgogne-Franche-Comté
Dijon, France
julien.dubois@u-bourgogne.fr

***Abstract*—This paper proposes a new navigation device to assist visually impaired people reach a defined destination safely in the indoor environment. This approach based on visual-auditory substitution provides the user a 2D spatial sound perception of the destination and of nearby and dangerous obstacles. Visual markers are placed at several relevant locations to create a mesh of the building where each marker is visually accessible from another marker. A graph representation of markers locations and their connection to each other defines by a way finding algorithm the shortest path reach to the wished position. The navigation task is achieved by moving from visual marker to visual marker until the desired destination is reached. These markers can be used independently of any other system or in addition to other solutions based on geolocalisation and/or a digital building model. Moreover, further information can be associated to the markers, and therefore verbalize to the user for instance a temporary hazards, a door presence or any other usual displayed information. The passive visual markers enables to deploy easily and quickly a scalable and low-cost solution to "signpost" the environment for users. Combined with our real-time implemented obstacle detection, their analysis enables the navigational abilities of visually impaired people to be improved.**

***Index Terms*—Auditory sensory substitution , Wearable assistive device, Navigation aid, Obstacle avoidance, Visual impairment, Sonification, Visual marker detection**

I. Introduction

A recent contribution to the WHO initiative, *VISION 2020: The Right to Sight* [1], estimate an increase of the number of people with visual impairment from 43.3 millions in 2020 to 61 millions in 2050 of blind people. This trend is worldwide and affects all groups of visually impaired people, from the low visually impairment to the totally blind. Indeed, this trend is caused by the increasing ageing and lack of treatment of visual degeneration such as glaucoma, presbyopia, etc. The increasing number of blind and severely visually impaired people, poses a challenge to enable them to lead an independent life while moving about safely. In fact, the variety of situations encountered in everyday navigation not only disrupts blind people's understanding of their destination but also, and above all, exposes them to risks such as the presence of a dangerous area such as a staircase, a static or mobile obstacle.

The ability to navigate in an environment is a fundamental element for an independent life. In daily life, a person is required to move frequently for various reasons such as work, leisure, etc. A person uses the spatial information acquired through their sensory modalities to orient themselves and move from one position to another. Although all sensory modalities are necessary to fully understand a scene, they do not provide the same level of spatial information. Vision is the most important sense for perceiving one's spatial environment. Indeed, it is easy to imagine how difficult it would be to navigate safely without visual perception in a familiar or unfamiliar environment.

A visually impaired person compensates by developing the spatial perception abilities of the other sensory modalities like the olfactory or the auditory sense, but some information is still not perceived. The degree of compensation varies according to the early age of onset of blindness. As the human body cannot fully compensate for this handicap, external means have been proposed to improve their navigation.

The classic methods used to remedy these problems are the white cane, the trained dog or the trained guide. Although these methods are widely used and effective for safe travel, they have limitations. The range of the white cane is limited to nearby objects, trained dogs and the guide require long and intensive training and are relatively expensive. In addition, due to the limitation of these means and the increase in the number of visually impaired people, the need for new technologies to enhance the spatial information perceived by a sighted person has increased. These devices or systems,

Thanks to the Conseil Régional de Bourgogne Franche-Comté, France and the Fond Européen de Développement Régional (FEDER) which are supporting financially this research.

called assistive technologies, must be ergonomic and have low latency to detect moving objects.

Assistive technologies are electronic systems designed to transmit information acquired by an artificial sensor to the blind person. In a review of different assistive technologies, [2] categorized devices according to the technologies used: vision replacement, vision enhancement, and vision substitution. Vision replacement and vision enhancement have some limitations. The first category relies on displaying information directly to the visual cortex of the brain and requires medical intervention, while the second category excludes people with severe visual impairment. Unlike other categories, Visual Substitution Devices (VSD) can be worn without limitation of use. In fact, the process transcribes visual information to another unimpaired sensory modality with out-of-body communication.

The auditory and tactile sensory modalities are primarily employed to transcode relevant information. Tactile-auditory substitution devices emit haptic feedback corresponding to a visual event while visual-auditory substitution devices emit a sound with verbal or non-verbal information. The VSD assist visually impaired persons in many tasks or situations of daily life, such as avoiding obstacles in the environment, defining the path to a desired location, or determining the precise position of the user. These three functionalities allow the devices to be grouped into three subcategories referred to in the literature as: Electronic Travel Assistance (ETA), Electronic Orientation Assistance (EOA) and Position Location Devices (PLD).

Fig. 1. Schematic view of the navigation aid system.

In this paper, we propose a new indoor EOA device based on a spatialized visual-auditory substitution approach. A navigation aid for visually impaired people provides information on the current trajectory, but above all, it must consider the presence or absence of obstacles that obstruct the user's progress. In addition, the system must be responsive with fast data processing and short sound emission to allow smooth movement to the desired destination. In response to this constraint, we proposed a system designed to be fast, easily integrated into an existing building, and easily expandable without economic cost. We propose a navigation method based on a mesh of the building by printable visual markers where the user navigates by intermediate beacons to reach the desired destination. The figure 1 shows a schematic view of the system operation in which a person receives a louder sound to the right indicating the relative position of the marker. The obstacle detection required for any navigation method is extracted from the depth map provided by the RGB-D camera. The 3D positions of the path and nearby obstacles are sonified into separate short 2D spatialized sounds. We propose to merge these separate information into a single sound in order to reduce the sound emission time, but also to have simultaneously the information on the trajectory and the spatial environment.

II. Related work

Navigational aids for blind people operate mainly by tracking the user's movements in real-time and guiding with feedback to the desired destination. However, the diversity of the visual environment or specific task space makes it difficult to establish a universal method of navigational aid. Indeed, some outdoor systems are based on the use of GPS (Global Positioning System) signals [3], [4], but the accuracy of GPS is limited and does not allow to precisely define the user's position in a building. Moreover, the problem is amplified in an indoor environment with the degradation of the GPS signal. Although GPS cannot be used to locate accurately in an indoor environment, research on location methods based on other types of waves and protocols has been performed [5]. Some used passive radio frequency identification (RFID) tag to guide the user from tag to tag until the final destination is reached [6] [7]. The predefined path is obtained by a path finding algorithm. The short operating range of these beacons requires a relatively large number of beacons to cover an entire space, which increases the cost of installation. Similar systems replace the passive RFID tag with an active tag or a Bluetooth beacon [8] [9]. However, despite the increased information flow and range of the beacons compared to passive devices. The use of active transmitters and receivers requires permanent battery power and therefore maintenance. Other competing systems use ultra-wideband (UWB) sensors to track the person in the environment [10]. This technology is more robust and the operating range of UWB (50-60m) allows it to be deployed in large spaces, but the unit cost of UWB transceivers is relatively higher than that of an RFID system for small spaces.

Similarly, the camera-based system [11] [12] is less complex to implement in any building with only an electronic device carried by the visually impaired person. Indeed, a visual marker detection evenly distributed in the spatial environment substitutes the use of electronic beacon. Detecting markers in visual space is resource and time-consuming and therefore disrupts the smoothness of navigation. In addition, a large distance between the user and a marker reduces the ability

to detect markers. [13] proposes a hybrid system with a visual marker and an active RFID is used to detect possible distant markers. However, the size of the marker is one of the main reasons for this difficulty, and the use of markers of varying size depending on the distance solves this problem.
The information, provided by the sensors, is enriched by semantic information obtained from the building information model (BIM) [14]. Indeed, a BIM representation combine with sensor data allow to represent the current position in the building space. Moreover, the BIM provides semantic information about the building configuration and the presence of hazards such as stairs and emergency exits.
Furthermore, knowledge of a fixed or mobile obstacle or more generally of a danger must be taken into account by the assisted navigation system. An obstacle will require a deviation from the initially predefined trajectory by bypassing it through a clear area. ETA devices are designed to identify obstacles around users. Some ETA methods guide the path to follow to circumvent the obstacle by indicating the area of free visual space via the subdivision of the space into distinct zones [15] or by using dynamic path methods inspired by the field of robotics based for example on an optimization of ant colonies [16]. Other methods just transmit information on the positions of obstacles [3] in order to allow a better understanding of the spatial environment and better freedom of movement.
Visual-auditory sensory substitution systems implement a protocol for transcoding the desired trajectory into an understandable sound. Most navigation aids provide verbal information to effectively guide the user [12] [17]. These approaches do not require special training for the handling of the system and allow a wide range of semantic description of the environment. In contrast, other methods propose to spatialize the information by emitting short spatialized stereophonic sounds to guide the user [18], [19]. This results in a shorter transmission time and therefore a lower latency. The ability to localize a static or moving object using spatialized stereophonic sounds has been studied.
The localization and sound generation processing is complex and requires an appropriate processing unit in order to obtain low latency information to ease navigation and the detection of possible dangerous zones. Approaches propose to perform processing via cloud computing [12] [17] in order to obtain a device with better ergonomics, performance and autonomy instead of an offline platform. Nevertheless, this type of device is subject to communication disturbances linked to the presence of a white spot in a building.
We propose a dynamic aid device with embedded processing, accurate in its indications, robust to disconnection, easily deployable and costless in all existing buildings. For this purpose, our proposed visual-auditory navigation aid system is offline, based on a detection of visual markers with an emission of short and spatialized 2D sounds.

III. Method

Our camera navigation system locates visual markers placed in the building. The markers are positioned at different points of interest such doors and in such a way another marker position is observable. In fact, the door crossing is an essential element for a useful navigation aid system in an indoor context. A spatial distribution of markers allows the user to move from marker to marker until he or she reaches the desired destination. Meshing the building with markers allows for a graphical representation of the environment with their associated connections and thus allows for the use of a graph-based pathfinding algorithm. However, the path finding algorithm requires that the nodes be identifiable and therefore the visual marker symbol used must be unique. The overall operation of our navigation's system is described in the figure 2 and below:

- **Initialization:** The path finding algorithm requires knowledge of the start node and the end node to initialize. The selection of the start node is performed by searching for a visual marker in the user's near space. After detecting the start node, the user is prompted for the desired destination.
- **User movement:** The user moves to the selected marker following a spatialized sound symbolizing the position in a sound space of the visual marker.
- **Marker reach:** When a marker is located within close enough to the user, the system checks if the marker is referenced as a point of interest and alerts the user if an action is required, such as opening a door. Then, if the marker reached is not the final destination, the user is prompted to scan the environment again to find the next marker.

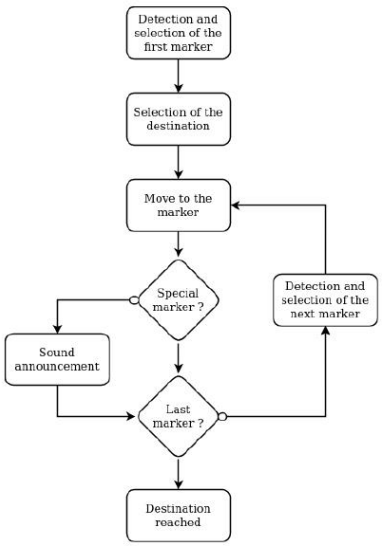


Fig. 2. Diagram of the navigation algorithm.

The detection and decoding of markers is a critical step in the operation of the system. The absence of detection prevents the user from continuing his journey, while the slow processing of visual information makes navigation less fluid or even jerky. Consequently, the choice of a marker with a unique and identifiable symbol as well as a robust detection method is necessary. We have chosen the STag fiduciary marker system [20] designed to be fast, stable, robust to detection of distant marker, to partial occlusion against difficult viewing angle conditions. Indeed, the detection of Stag markers is combined on an extraction of geometric features such as ellipse, corner and edge with a refinement of the homography to allow a real-time processing. In addition, the large number of individual markers in the libraries allows deployment in large indoor spaces. An example of three separate STags is provided in figure 3.

Fig. 3. Example of different STag visual markers [20]

Our undirected graph representation of the building mesh consists of nodes that are linked to each other without knowing the distance between them. An unweighted graph allows us to use a Depth First Search (DFS) or Breadth-First Search (BFS) brute force path finding algorithm to obtain the path to the destination. These algorithms are similar in their method with the search for the first occurrence of the arrival node by traversing the graph from the departure node. These algorithms are similar in their method with a graph traversal from the start node to the first occurrence of the end node. The main distinction is the traversal mode: DFS traverses the graph in depth while BFS traverses the graph in width. Although DFS is quicker and needs less memory than BFS to obtain a solution, the obtained solution is suboptimal in contrast to BFS. In our method, we have privileged the BFS algorithm in order to propose to the visually impaired person the shortest path to reach the destination.

However, a linear navigation is impossible in complex environment with various multiple static or moving obstacles such as an indoor space. We have integrated in our EOA method informations about nearby obstacle of the user in order to increase the scene understanding and to avoid the obstacle. Moreover, a better scene understanding allow more freedom of movement in the navigation task. Nearby obstacles are extracted from the depth video stream of the RGB-D camera with a threshold of elements located in the visual scene within one meter of the user.

Our sonification method is based on a low latency auditory substitution approach [21] where each visual position is associated with a short spatialized stereophonic audio pixel. A spatialize sound allows a precise localization of an area of interest without emitting a long verbal expression which increases the delay between two successive information and thus increases the danger of navigation. In the interest of limiting the computational delay, we have pre-computed for each pixel position the corresponding audio pixel. The spatialization of a given pixel into an audio pixel is based on the convolution of a monophonic sound with the impulse response associated with the corresponding azimuthal position in the head-related impulse response (HRIR) data set. A HRIR is a impulse response with mimic natural deformation made by body of the listener mainly the head on the sound to know its origin. The coordinate of the pixel is mapped into a spherical coordinate according to expression (1). With *fov* the azimuthal field of view of the camera, *pos* the position of the pixel and *size* the width of the image in pixel.

$$angle = \frac{2 * pos - size}{size * fov} \tag{1}$$

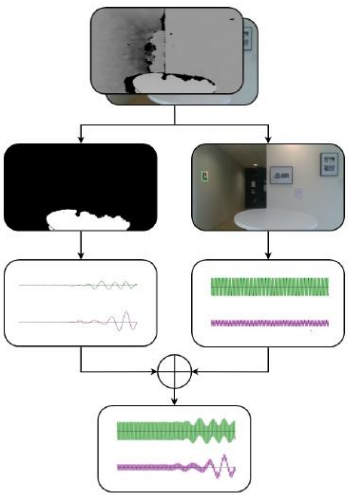


Fig. 4. Video processing and sonification pipeline for obstacle detection (left column) and marker detection (right column). The green and purple colours symbolise the left and right channels of stereo sound respectively.

The visual data is a combination of path and obstacle information. However, both information must be easily understandable and differentiable by the user. In addition, all information must be provided to the user at the same time, without alternation, to enable safe and smooth navigation. The combination of visual information into a unique understandable auditory signal is the cornerstone of our EOA method. To each of these visuals information, we associate a unique and brief monophonic sound, easily identifiable. Then we spatialized these sounds using a two-metre HRIR. Finally, we add the resulting sounds together to obtain an overall audio sound. The figure 4 summarizes

our visual processing and sonification method. The obstacle detection (left) and the path direction (right) information are sonified into two distinct spatialized sounds. The green colour and the purple colour symbolise respectively the left and right channel of the stereophonics sound. Obstacle detection (left) and path direction (right) information are sonified into two distinct spatialized sounds. The green and purple colours symbolize the left and right channels of a stereophonic sound. The last graph represents the sum of these sounds in an overall sound.The difference in amplitude between the left and right channels on the navigation sound indicates that a marker is located on the right. On the obstacle sound, the left and right channels are similar and symbolize an obstacle in the centre and at the bottom. Two distinct monophonic sounds are used to spatialize and distinguish auditory information. The obstacle monophonic sound have a lower frequency than the navigation monophonic sound.

In addition to these spatialized sounds and due to the multiple steps to reach the desired destination. We have added verbal information about the current stage but also information if a door has to be passed or a lift has to be used. Semantic sounds are pre-recorded and are played at the beginning or end of a step.

IV. System

The pipeline of our system is similar to that of other sensory substitution devices, with an acquisition step performed by a camera, followed by processing of the visual data by a computational unit, and then transcription and output into auditory information by an audio device. The videos are acquired by an Intel RealSense D435i stereoscopic camera with a depth field of view (FOV) of 87° × 58° and a colour camera FOV of 69° × 42°. The colour and depth image were synchronously captured and realigned with a resolution of 1280 × 720 pixels at 30 frames per second. The camera module is placed at eye level on electronic glasses to allow the blind person to scan the scene with a head movement. The processing unit used is a standard laptop in a backpack with the Ubuntu 20.04 operating system, an Intel Core i7-6700HQ processor (4 Cores - 8 Threads with a frequency of 2.60 GHz), 8 GB of RAM, and a Nvidia GTX 1070 mobile graphics card. Bluetooth's headphones are used to transmit auditory information.

The software is developed in C++ using the LibRealsense2 , OpenCV libraries for video acquisition and processing and Advanced Linux Sound Architecture libraries (ALSA) for writing the sound driver. Visual marker detection is performed on a colour image converted to grey scale required by the STag detection algorithm, and with the highest possible input resolution provided by the camera in order to identify the marker pattern over a large distance. The audio pixel sonification of nearby objects is performed at a resolution of 160 x 120 pixels. The resolution is limited by the HRIR dataset used to spatialize a sound, but mostly by the inability of the human ear to distinguish a small angular variation of the sound emission.

A realistic daily navigation aid system for visually impaired people must meet response time constraints. Indeed, a significant processing delay between the acquisition and the associated sound emission can disturb the user, or even cause serious injuries if the information of an obstacle is not received in time by the user. The limitation of delays is therefore a necessity, for which we have optimized our method with preloaded sound and multi-threaded approaches. The program is split into 4 threads where each one performs the following function:

- **Acquisition thread:** Acquisition of the depth and colour image, the alignment the depth map to the colour image.
- **Video processing thread:** Detection and localization of the STag markers in the RGB image and of the nearby objects on the depth map.
- **Manager thread:** Based on our navigation method, this thread generates an audio sound according to a visual information given by the video processing thread or a verbal sound expression.
- **Sound thread:** Loading the generated sound or generated speech to be transmitted to the user according to the navigation manager.

Our multithreaded approach shown in figure 5 is based on the sequential operation of sensory substitution systems. The navigation manager is the central node of the system that links the video processing stream and the audio output stream. The separation of the audio and video processes reduces the system delay, so that while the audio is being transmitted, the next visual information is already being processed. In addition, this separation allows continuous sound to be output without disturbing the user with an absence of sound. A lack of sound may be due to a slowdown in the video stream or to images being lost during acquisition. The sound is emitted in such a way that if new visual information has not yet arrived, then information from nearby objects is re-emitted.

The table I summarises the impact of the different video and audio processes during a navigation between two markers. The processing time for the sonification of nearby objects and the detection of STag markers varies slightly with the number of nearby elements or markers in the visual scene. Our measurement of processing times was realized in a scene with a marker displayed on the wall at 60 cm. The detection time of an environment is given for the detection of a single STag marker, which represents the general use case of the system with markers evenly distributed in the spatial environment. The overall system time of approximately 78 ms allows satisfactory navigation in an indoor space. Furthermore, our EOA is a CPU-only implementation and does not take advantage of the fact that most of the algorithms used are sequential and allow GPU optimization. An implementation of the segment detection algorithm used for marker detection with CUDA shows a speed-up of $\times 12$ [22] and an overall detection speed-up estimated at $\times 4.4$ on older GPUs [20]. GPU optimization limits CPU usage, reduces overall system time, and could enable real-time hardware implementation on a low-power

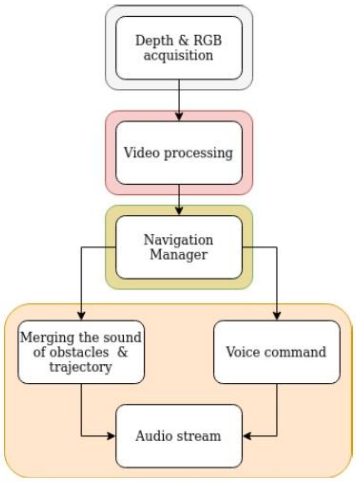


Fig. 5. Implementation's diagram

embedded target with a GPU module such as the Nvidia Jetson cards.

TABLE I
SYSTEM'S PROCESSING TIME ON THE LAPTOP TARGET (TIME IN MILLISECOND)

Video processing		Sonification		
STag	Nearby Obs.	STag	Nearby Obs.	Sound fusion
57	1.4	0.02	17.5	1.9

V. EXPERIMENTATION

We measured the capabilities offered by our navigation aid system with a short sound emission combining a path direction and an obstacle information. As a proof-of-concept, we asked a blindfolded person with no specific training with a visual-auditory substitution device to reach an unknown destination with our system within a marked building. The destination is entered by a supervisor without the user being informed of this information. Before the start of the experiment, a brief explanation of the system and of the two sound information emitted was given. For this purpose, we created a realistic scene within a marked building with obstacles placed. Indeed, a realistic experimentation of an indoor navigation system requires an environment that represents the daily life of visually impaired people.

The experimental space is a part of the 3rd floor of the I3M building (ImViA & LEAD Laboratories, Dijon). This space contains several static obstacles (chairs, desks, tables, ...) with 6 markers printed on a A4 paper and placed at relevant position in order to create a mesh the building. The figure 6 shows the graph representation of the mesh used to compute the way

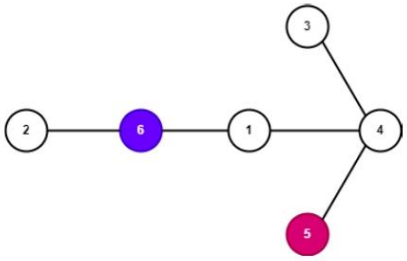


Fig. 6. Graph representation of the indoor space used for the experimentation. The possible navigation between two nodes is symbolized by a link. The purple node represents a close door and the pink node a lift.

finding algorithm to reach the desired destination. A door is indicated with a purple node.

The starting position of the blindfolded person is aligned with the marker 2 in the room and the unknown position is the floor's lift indicated by the marker 5. The path predefined by the way finding algorithm to reach this destination from the initial position is a transit through the clues $2 \rightarrow 6 \rightarrow 1 \rightarrow 4 \rightarrow 5$. The user had to pass through an open door to reach marker 2 and open the door indicated by marker 6 to reach marker 1. Finally, it had to avoid various static obstacles positioned between markers 1 - 4 and 4 - 5. Figure 7 represents a top view of the floor building captured by multiple Lidar (laser imaging detection and ranging) scans with information about the positions of the visual markers (green and purple arrows), the participant's starting (red dot) and destination (pink dot) positions, and the path followed (white line).

The path taken to reach marker 5 was recorded in order to analyse the viability of our indoor navigation aid system and specifically the user's behaviour when encountering an obstacle. The relative position of the person in the building was recorded using an Oculus Quest 2 headset in parallel with our system. The oscillations of the trajectory in the path followed by the user are due to the swaying of the human being when walking but also to the movements of the user's head necessary to scan the environment and possible obstacles. Indeed, the inertial measurement unit sensor (IMU) is positioned in front of the user's eyes (Oculus Quest 2 computing unit) and is therefore sensitive to head movements. The maximum distance between the current marker and the person before moving to the next marker is 80 cm. The path followed by the blindfolded person represents the difficulties of the route. Indeed, the person backs up slightly to locate himself before opening the door, then detects and avoids the tables to reach markers 4 and 5. The experiment can be successfully reproduced and the results shows that a user could navigate in an outdoor environment by avoiding static obstacles (including the wall) or mobile obstacles (i.e anyone walking past) and passing through a door to reach the desired destination or passing through a door to reach the desired destination using our visual to auditory sensory substitution device. In addition,

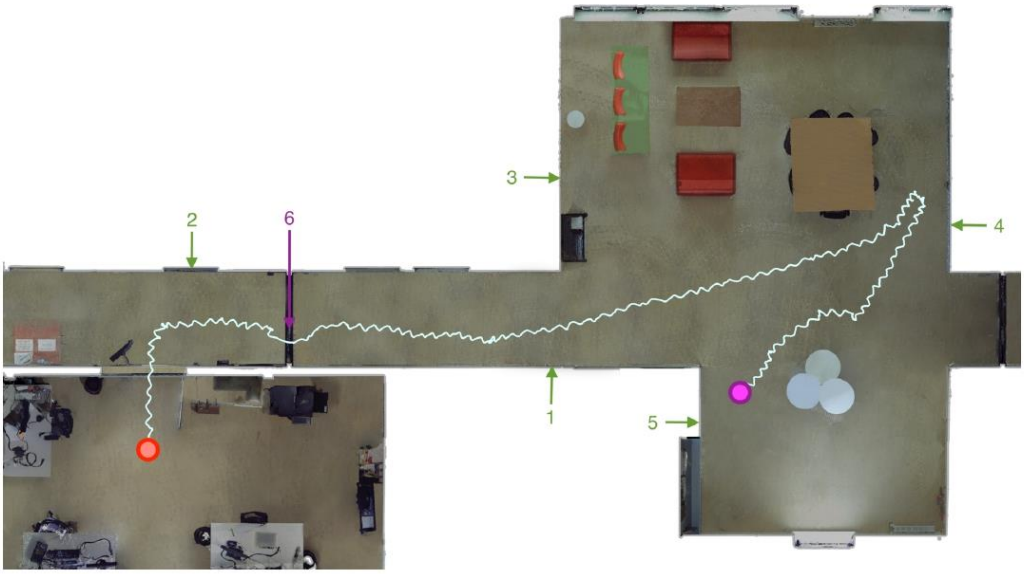


Fig. 7. Top view of the user's movements to reach the desired destination. The red dot and pink points indicate the starting position and the destination respectively. The arrows represent the position of the markers in the building space and the purple arrow indicates that the marker is placed on a door. The number associated to an arrow represents the tag number.

additional information can be included in any visual marker to be verbalized.

VI. Conclusion & Future work

Our proposed indoor navigation system allows a blind person to reach a desired destination in a building by guiding 2D spatialized sounds based on the recognition of unique visual markers. In addition to trajectory information, the presence of nearby obstacles or key points, such as a door, is provided. Our method can be adapted to any building or room and does not require remote transmission to achieve real-time performance. However, this system have some limitations. Indeed, the detection distance of the marker is limited in terms of distance, and moreover is sensitive to marker occlusion by dynamic obstacles. In addition, the position of the camera at the eye level cause a issue with the detection of small obstacle on the floor. In fact, reason is the height between the feet and the eyes combined with a low vertical camera's FOV makes it impossible to detect a obstacle. This problem could be solved by looking down from time-to-time to check whether an object is present or not. The use of a conventional white cane in addition to our method could also solve this problem. In fact, a white cane is the opposite of our obstacle detection with an excellent low obstacle but ineffective for high obstacles. The design of our EOA is not ergonomic for the user, with excessive weight and power consumption due to the use of a laptop as a computing unit. An optimization of our method on a low-power device with a GPU implementation is possible. Furthermore, the obtrusive nature of the ear canal with headphones could interfere with the understanding of the natural auditory scene. However, a replacement with bone conduction earphones avoids the auditory obstruction. The navigation method can be enhanced with an integration of voice command processing to interact with the system in order to obtain some additional information about the spatial scene. The semantic information could be a brief description of the person's location, or a description of the type of obstacle on the path if the user requests it. Moreover, a fusion with the building information model could provide information on the spatial organisation of the building and the presence of hazardous areas such as stairs. Finally, a cognitive psychology study could be considered to investigate the user's behaviour with our sonification method.

References

[1] Bourne and al., "Trends in prevalence of blindness and distance and near vision impairment over 30 years: an analysis for the Global Burden of Disease Study," *The Lancet Global Health*, vol. 9, no. 2, pp. 130–143, Feb. 2021.

[2] W. Elmannai and K. Elleithy, "Sensor-Based Assistive Devices for Visually-Impaired People: Current Status, Challenges, and Future Directions," *Sensors*, vol. 17, no. 3, p. 565, Mar. 2017.

[3] M. Poggi and S. Mattoccia, "A wearable mobility aid for the visually impaired based on embedded 3D vision and deep learning," in *IEEE Symposium on Computers and Communication (ISCC)*. Messina, Italy: IEEE, Jun. 2016, pp. 208–213.

[4] A. Brilhault, S. Kammoun, O. Gutierrez, P. Truillet, and C. Jouffrais, "Fusion of Artificial Vision and GPS to Improve Blind Pedestrian Positioning," in *IFIP International Conference on New Technologies, Mobility and Security*. Paris: IEEE, Feb 2011, pp. 1–5.

[5] J. Xiao, Z. Zhou, Y. Yi, and L. M. Ni, "A Survey on Wireless Indoor Localization from the Device Perspective," *ACM Computing Surveys*, vol. 49, no. 2, pp. 1–31, Nov. 2016.

[6] C. Tsirmpas, A. Rompas, O. Fokou, and D. Koutsouris, "An indoor navigation system for visually impaired and elderly people based on Radio Frequency Identification (RFID)," *Information Sciences*, vol. 320, pp. 288–305, Nov. 2015.

[7] R. Ivanov, "Indoor navigation system for visually impaired," in *International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing on International Conference on Computer Systems and Technologies - CompSysTech '10*. Sofia, Bulgaria: ACM Press, 2010, p. 143.

[8] S. A. Cheraghi, V. Namboodiri, and L. Walker, "GuideBeacon: Beacon-based indoor wayfinding for the blind, visually impaired, and disoriented," in *IEEE International Conference on Pervasive Computing and Communications (PerCom)*. Kona, Big Island, HI, USA: IEEE, Mar. 2017, pp. 121–130.

[9] D. Ahmetovic, C. Gleason, C. Ruan, K. Kitani, H. Takagi, and C. Asakawa, "NavCog: a navigational cognitive assistant for the blind," in *International Conference on Human-Computer Interaction with Mobile Devices and Services*. Florence Italy: ACM, Sep. 2016, pp. 90–99.

[10] A. Martinez-Sala, F. Losilla, J. Sánchez-Aarnoutse, and J. García-Haro, "Design, Implementation and Evaluation of an Indoor Navigation System for Visually Impaired People," *Sensors*, vol. 15, no. 12, pp. 32 168–32 187, Dec. 2015.

[11] G. E. Legge, P. J. Beckmann, B. S. Tjan, G. Havey, K. Kramer, D. Rolkosky, R. Gage, M. Chen, S. Puchakayala, and A. Rangarajan, "Indoor Navigation by People with Visual Impairment Using a Digital Sign System," *PLoS ONE*, vol. 8, no. 10, p. 76783, Oct. 2013.

[12] Y.-J. Chang, S.-K. Tsai, and T.-Y. Wang, "A context aware handheld wayfinding system for individuals with cognitive impairments," in *International ACM SIGACCESS conference on Computers and accessibility - Assets '08*. Halifax, Nova Scotia, Canada: ACM Press, 2008, p. 27.

[13] S. Alghamdi, R. van Schyndel, and A. Alahmadi, "Indoor navigational aid using active RFID and QR-code for sighted and blind people," in *IEEE International Conference on Intelligent Sensors, Sensor Networks and Information Processing*. Melbourne, VIC: IEEE, Apr. 2013, pp. 18–22.

[14] R. Marroquin, J. Dubois, and C. Nicolle, "Ontology for a panoptes building: Exploiting contextual information and a smart camera network," in *Semantic Web*, Aug 2018, vol. 9, p. 803 – 828.

[15] W. M. Elmannai and K. M. Elleithy, "A highly accurate and reliable data fusion framework for guiding the visually impaired," *IEEE Access*, vol. 6, pp. 33 029–33 054, Mar 2018.

[16] A. Benabid Najjar, A. Rashed Al-Issa, and M. Hosny, "Dynamic indoor path planning for the visually impaired," *Journal of King Saud University - Computer and Information Sciences*, p. S1319157822000751, Mar. 2022.

[17] J. Bai, D. Liu, G. Su, and Z. Fu, "A Cloud and Vision-based Navigation System Used for Blind People," in *Proceedings of the 2017 International Conference on Artificial Intelligence, Automation and Control Technologies - AIACT '17*. Wuhan, China: ACM Press, 2017, pp. 1–6.

[18] M. J. Proulx, P. Stoerig, E. Ludowig, and I. Knoll, "Seeing 'where' through the ears: Effects of learning-by-doing and long-term sensory deprivation on localization based on image-to-sound substitution," *PLOS ONE*, vol. 3, no. 3, pp. 1–8, Mar 2008.

[19] F. Scalvini, C. Bordeau, M. Ambard, C. Migniot, and J. Dubois, "Low-latency human-computer auditory interface based on real-time vision analysis," in *ICASSP - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 36–40.

[20] B. Benligiray, C. Topal, and C. Akinlar, "Stag: A stable fiducial marker system," *Image and Vision Computing*, vol. 89, pp. 158–169, Sep 2019.

[21] M. Ambard, "Software Design for Low-Latency Visuo-Auditory Sensory Substitution on Mobile Devices," *Computer and Information Science*, vol. 10, no. 2, p. 1, 2017.

[22] O. Ozsen, C. Topal, and C. Akinlar, "Parallelizing edge drawing algorithm on CUDA," in *IEEE International Conference on Emerging Signal Processing Applications*. Las Vegas, NV: IEEE, Jan. 2012, pp. 79–82.

Résumé

Titre : Développement d'un dispositif de substitution sensorielle vision-vers-audition : étude des performances de localisation et comparaison de schémas d'encodage

Mots-clés : substitution sensorielle, perception spatiale auditive, environnement virtuel, analyse de scènes auditives, localisation auditive, déficience visuelle

Les dispositifs de substitution sensorielle vision-vers-audition convertissent des informations visuelles en un paysage sonore dans le but de permettre de percevoir l'environnement à travers la modalité auditive lorsque la modalité visuelle est altérée. Ils constituent une solution prometteuse pour améliorer l'autonomie des personnes déficientes visuelles lors de leurs déplacements pédestres. Ce travail de thèse avait pour objectif principal de déterminer et d'évaluer un schéma d'encodage pour la substitution sensorielle permettant la perception spatiale 3-dimensionnelle en proposant des protocoles de familiarisation et d'évaluation dans des environnements virtuels plus ou moins complexes. Le premier objectif était de déterminer si la reproduction d'indices acoustiques pour la perception spatiale auditive était plus efficace que l'utilisation d'autres indices acoustiques impliqués dans des interactions audio-visuelles. La première étude a mis en évidence que la modulation de la hauteur tonale dans le schéma d'encodage permettait de compenser en partie les limites perceptives de la spatialisation pour la dimension de l'élévation. La deuxième étude a mis en évidence que la modification de l'enveloppe sonore pouvait permettre de compenser la perception compressée de la distance. Le deuxième objectif de ce travail de thèse était de déterminer dans quelle mesure le schéma d'encodage utilisé préservait les capacités de perception spatiale dans un environnement complexe composé de plusieurs objets. La troisième étude a mis en évidence que les capacités de ségrégation d'une scène visuelle complexe à travers le paysage sonore associé dépendaient de la signature spectrale spécifique des objets la composant lorsque la modulation de la hauteur tonale est utilisée comme indice acoustique dans le schéma d'encodage. Les travaux de cette thèse ont des implications pratiques pour l'amélioration des dispositifs de substitution concernant, d'une part, la possibilité de compenser les limites perceptives spatiales avec des indices acoustiques non-spatiaux dans le schéma d'encodage, et d'une autre part, la nécessité de réduire le flux d'informations auditives pour préserver les capacités de ségrégation du paysage sonore. Les protocoles de familiarisation et d'évaluation en environnement virtuel ayant été développés de sorte à être adaptés à la population déficiente visuelle, les travaux de cette thèse soulignent le potentiel des environnements virtuels pour évaluer précisément les capacités d'utilisation de dispositifs de substitution dans un contexte contrôlé et sécurisé.

Abstract

Title: Development of a visual-to-auditory sensory substitution device: study of localization performance and comparison of encoding schemes

Keywords: sensory substitution, auditory spatial perception, virtual environment, auditory scene analysis, auditory localization, visual impairment

Visual-to-auditory sensory substitution devices convert visual information into soundscapes for the purpose of allowing the perception of the environment with the auditory modality when the visual modality is impaired. They constitute a promising solution for improving the autonomy of visually impaired people when traveling on foot. The main objective of this thesis work was to determine an encoding scheme for sensory substitution allowing 3-dimensional spatial perception by proposing familiarization and evaluation protocols in virtual environments with different complexities. The first aim was to determine whether the reproduction of acoustic cues for auditory spatial perception was more effective than the use of acoustic cues involved in audio-visual interactions. The first study demonstrated that the modulation of pitch in the encoding scheme could partly compensate for the perceptual limits of spatialization for the dimension of elevation. The second study showed that the modification of the sound envelope could partly compensate for the compressed perception of distance. The second objective was to determine to what extent the determined encoding scheme preserved spatial perception abilities in a complex environment where several objects were present. The third study demonstrated that the segregation capabilities of a complex visual scene through the soundscape depend on the specific spectral signature of the objects composing it when pitch modulation is used as an acoustic cue in the encoding scheme. The work of this thesis has practical implications for the improvement of substitution devices concerning, on the one hand, the possibility of compensating spatial perceptual limits with non-spatial acoustic cues in the encoding scheme, and on the other hand, the need to reduce the amount of auditory information to preserve the segregation abilities of the soundscape. The familiarization and evaluation protocols in a virtual environment having been developed to be adapted to the visually impaired population, the work of this thesis highlights the potential of virtual environments to precisely evaluate the abilities to use sensory substitution devices in a secure context.